Sweeping Preconditioners for Helmholtz Problems using Absorption

Arter, Elizabeth

Award date:
2019

Awarding institution:
University of Bath

Sweeping Preconditioners for Helmholtz Problems using Absorption

submitted by

Elizabeth Anne Arter

for the degree of Doctor of Philosophy

University of Bath

Department of Mathematical Sciences

November 2018

COPYRIGHT

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

.....

Elizabeth Arter

Signed on behalf of the Faculty of Science .....

Summary

The discretisation of boundary value problems for the Helmholtz equation (frequency domain wave equation) leads to linear systems that are non-self adjoint and highly indefinite. The iterative solution of these problems is difficult and much recent research has focussed on the construction of good preconditioners. This thesis examines the effect of adding absorption to sweeping-type preconditioners for Helmholtz problems. The application of interest for the Helmholtz problems in this thesis is seismic inversion.

First we look at the beneficial effects of absorption on low-rank separable expansions for the Hankel function, that provide valuable theoretical motivation for sweeping-type preconditioners. We find that when absorption is included, there are three ways in which benefits are seen: the quality of the separable expansion increases, or the size of the domains for which the separable expansion is valid increases, or a lower rank may be sufficient to gain the same quality of separable expansion.

Next we focus on the effect of adding absorption to Schur complement matrices arising in the construction of sweeping-type preconditioners. The theoretical and numerical results show good agreement on the following points: the dependence of the rank upon the quality of the approximation, the independence of the rank from the wavenumber, the exponential improvement in the quality of the approximation with absorption and the ranks remaining low for taller domains when absorption is included.

Finally we look at the effect of adding absorption on the iteration counts of several variants of sweeping preconditioners. We find that in some cases we see improvements due to absorption and in others we do not. The performance of the iterative method is highly dependent on the parameters used in both the discretisation of the problem and the construction of the preconditioners.

Acknowledgements

Firstly, I wish to thank my supervisors Euan Spence and Ivan Graham for their invaluable support throughout my Ph.D., for the productive mathematical discussions, as well as their attention, guidance, patience and giving feedback on the written drafts of this thesis. I also wish to thank Adrian Hill and Alison Ramage for their suggested improvements to this thesis during the examination process and also for Adrian's support and diligence as my Director of Studies.

I am grateful to Paul Childs, James Rickett and Can Evren Yarman for interesting conversations and opportunities to learn more about the applications of the work conducted in this thesis. I am also grateful to the organisers and attendees of the Numerical Analysis Seminar and WAVES conferences, where I learned much of interest about the wider mathematics around my project.

I am highly indebted to Lexing Ying for providing me with copies of code for the Engquist and Ying sweeping preconditioner and Stephanie Meier-Rohr for providing me with copies of her code for doing calculations with strongly admissible $\mathcal{H}$-matrices, both used in numerical experiments in this thesis.

I would like to thank those in the university support services, without whom this thesis would not have been typed, especially Petra Harwin, Emma Cliffe and Susan Fielding. I would also like to thank Josh, for his work typing this thesis.

I wish to thank the Engineering and Physical Sciences Research Council and Schlumberger Cambridge Research for their funding of this project through my CASE studentship.

I wish to acknowledge my fellow Ph.D. students for their friendship, advice and the many other ways they have been of invaluable help to me, especially those of 4W1.15, Wednesday cake attendees, Drinks in the Parade attendees and the board games group. Special thanks to Siân, Matt, Andrew, Mason, Steven, Kieran, Qian, Alge, Will, Sam, Jack, Cameron, Beth, Owen, Dan and Josh.

My deepest thanks go to my Mum and Dad, Annette and Wayne Arter, for always being there for me during my Ph.D. as my loving and supportive parents. Also to wider friends and family too numerous to list, for their sustaining friendship.

I wish to thank all the members of Emmanuel Church Bath, including Ad, Jane, Jonny, Clare, Tom, Katy, Dom, Julia, Chris, Serena, Carl, Bex, Hutch,

Lynds, James, Margaret, Rupert and most especially Margie, for fellowship, spiritual and physical food and prayers. I also wish to thank Kate Ellis-Sawyer, Amira Battle, housemates including Sophie, Karisha and Chloe, the Christian Postgraduate Group including Heather, Eleanor, Joel, Elizabeth and Lewis and the Emmanuel Church Bath students, including Luke, Mathew, Eivind, Jemima, Chelsea, Esther, Henry, Josiah, Jackson, Isobel, Simon and Sam for their fellowship, friendship and prayers.

Lastly, I wish to thank God. It is impossible to expound all the ways I am indebted to him, as they include his creation of all things (which are too numerous to list), though of special relevance are the mathematics in this thesis, me and the people listed above. Most important is his planning of the central event of history, where Jesus Christ loved and gave himself for the world[1], as the propitiation for sins[2] and reconciling even me to God[3], so that I have known him and his most precious and ever-present love, encouragement and help throughout my Ph.D..

Soli Deo gloria
Glory to God alone

[1]Galatians 2:20 and John 3:16 Holy Bible ESV [1]
[2]1 John 2:2 Holy Bible ESV [1]
[3]2 Corinthians 5:18 Holy Bible ESV [1]

Contents

List of Figures

List of Tables

Chapter 1

Introduction: Preconditioners for Helmholtz problems

1.1 Helmholtz Problems

1.1.1 The Helmholtz Equation

This thesis considers solvers for discretisations of Helmholtz-equation problems, especially in the high-frequency regime. To establish the setting for these problems, we first give some background into the Helmholtz equation itself.

The Helmholtz equation is a time-independent partial differential equation which is the simplest possible model of wave propagation. Indeed, observe that the wave equation

$$\frac{\partial^2}{\partial t^2}U(x,t) - c^2(x)\Delta_x U(x,t) = F(x,t), \tag{1.1}$$

where Δ_x is the Laplacian with respect to x, with time-harmonic solution $U(x,t) = u(x)\exp(-\mathrm{i}\omega t)$ and time-harmonic source $F(x,t) = f(x)c^2(x)\exp(-\mathrm{i}\omega t)$, reduces to the Helmholtz equation

$$\left(\Delta_x + k^2(x)\right)u(x) = f(x), \tag{1.2}$$

where $k(x) := \omega/c(x)$ is the wavenumber, $c(x)$ is the wavespeed and ω the angular frequency [21, §3]. Alternatively, the Helmholtz equation arises from taking the

Fourier transform in time of the wave equation (1.1). As $k \to 0$, the solutions become less oscillatory and more 'Laplace-like'. As k increases, the solutions become more oscillatory. For example, for the homogeneous Helmholtz equation in 1D, with $x \in \mathbb{R}$, the general solution is given by $u(x) = A\sin(kx) + B\cos(kx)$. As k increases, Helmholtz problems become harder to solve; indeed the limit $k \to \infty$ is a singular limit, as the coefficient of the highest-order term in (1.2) vanishes. In this thesis we are concerned with problems in the mid- to high-frequency range.

1.1.1.1 Fundamental Solutions

Definition 1.1.1. ***(Fundamental Solution*** Φ***)*** *A fundamental solution* Φ *of a linear partial differential operator* $\mathcal{L}$ *on* $\mathbb{R}^d$*, is defined to be a function such that*

$$\mathcal{L}_x(\Phi(x,y)) = -\delta(y-x) \text{ in the distributional sense, for all } x, y \in \mathbb{R}^d,$$

where δ *is the Dirac delta function.*

Note that the fundamental solution is also known as the free-space Green's function.

The solution of the problem $\mathcal{L}(v) = -g$ on $\mathbb{R}^d$, where g has compact support, is then given in terms of the fundamental solution as follows:

$$v(x) = \int_{\text{supp}(g)} \Phi(x,y)g(y)\,\mathrm{d}y.$$

For the Helmholtz operator with constant wavenumber k, we have $\mathcal{L} = \Delta + k^2$, and expressions for the fundamental solutions are known. We choose the fundamental solutions that satisfy the outgoing radiation condition (discussion about the radiation condition is in §1.1.2 below).

In 2D the fundamental solution of the Helmholtz operator can be expressed in terms of the Hankel function of the first kind (in much of the rest of the thesis the superscript (1) will be dropped for notational convenience):

$$\Phi(x,y) = \frac{\mathrm{i}}{4}H_0^{(1)}(k\|x-y\|), \quad x, y \in \mathbb{R}^2, \tag{1.3}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^2$.

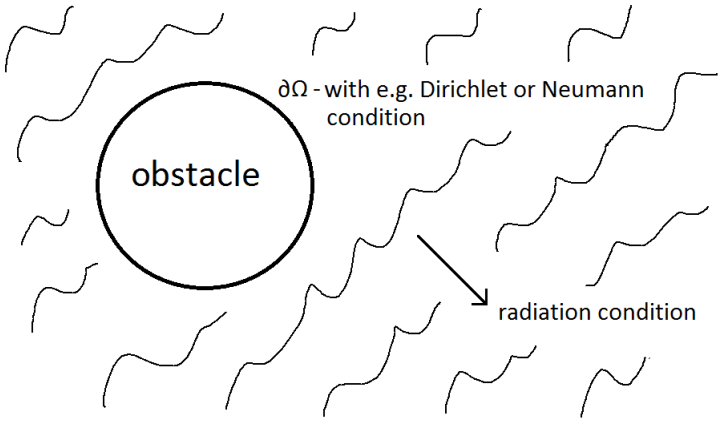


Figure 1-1: Sketch of scattering problem with radiation condition.

In 3D the fundamental solution of the Helmholtz operator is

$$\Phi(x,y) = \frac{\exp(ik\|x-y\|)}{4\pi\|x-y\|}, \quad x,y \in \mathbb{R}^3, \tag{1.4}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^3$.

More details about fundamental solutions can be found in [21, §3].

1.1.2 Boundary Value Problems involving the Helmholtz Equation

In this thesis, we only consider scattering boundary value problems (BVPs) posed on unbounded domains. There are many applications for such wave scattering problems and methods for their efficient solution are an area of active research.

A classical scattering problem involves wave scattering by a surface or an obstacle and the waves are allowed to propagate on an unbounded domain. There are many possible "obstacles", for example impenetrable objects with Dirichlet or Neumann boundary conditions, called "sound-soft" or "sound-hard" respectively, see Figure 1-1. Due to our application of interest (seismic inversion, see §1.1.3.1), in this thesis we only consider scattering by a "penetrable object", modelled by using the Helmholtz operator with spatially-varying wavespeed $c(x)$, resulting in spatially-varying wavenumber $k(x) := \omega/c(x)$. We further assume that both the source f and the variation in the wavespeed have support that is confined within

some bounded domain of interest.

For such scattering problems posed on unbounded domains, we need an additional condition at ∞ to ensure the problem has a unique solution. For this we use the Sommerfeld radiation condition (SRC), see Definition 1.1.2. The SRC physically corresponds to imposing the condition that the waves are outgoing and decaying towards infinity, originating from sources of energy within the finite area of interest.

Definition 1.1.2. ***(Model Problem)*** *Let the wavespeed $c(x)$ be a spatially varying function such that $(1 - c(x))$ has compact support and let $f(x)$ have compact support. Let $u(x)$ be the solution of the Helmholtz equation*

$$\Delta_x u(x) + k^2(x)u(x) = -f(x), \quad x \in \mathbb{R}^2, \tag{1.5}$$

with $k(x) := \omega/c(x)$, with the Sommerfeld radiation condition (SRC)

$$\frac{x}{\|x\|} \cdot \nabla u(x) - \mathrm{i}k(x)u(x) = o\left(\frac{1}{\|x\|}\right) \text{ as } \|x\| \to \infty. \tag{1.6}$$

Remark 1.1.3. *Observe that as $(1-c(x))$ has compact support, $k(x)$ is constant outside a compact set and therefore k is constant in* (1.6).

This problem has a unique solution [22, Theorem 8.7].

Remark 1.1.4. *We highlight here that it can be shown that the Sommerfeld radiation condition is not a self-adjoint boundary-condition operator, see for example [105, Example 4.9], ultimately due to the presence of the complex number* i. *So we note here that scattering problems with the Sommerfeld radiation condition are not self-adjoint.*

1.1.2.1 Green's Functions

We recall the definition of the Green's function.

Definition 1.1.5. ***(Green's function G)*** *Let $\mathcal{L}$ be a linear partial differential operator, let Ω be a bounded domain and consider the BVP $\mathcal{L}v = g$ in Ω, with some boundary conditions $Bv = 0$ on $\partial\Omega$. Then a Green's function for this BVP*

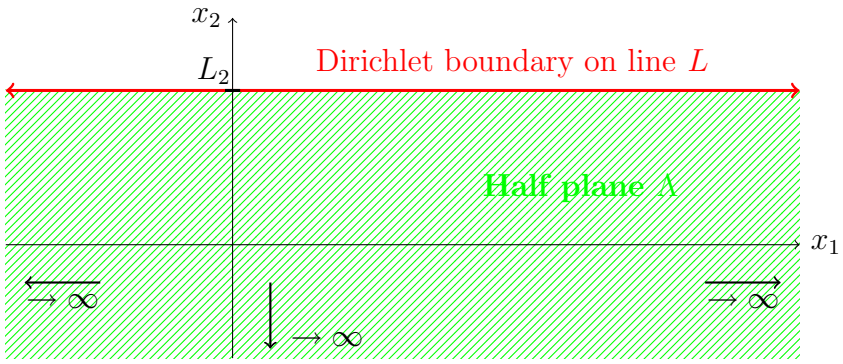


Figure 1-2: Half-plane domain.

is a function that satisfies

$$\mathcal{L}_x(G(x,y)) = -\delta(y-x) \text{ in the distributional sense, for all } x, y \in \Omega$$

and

$$B(G(x,y)) = 0, \text{ for all } y \in \partial\Omega$$

where B is applied with respect to the variable x.

In the case of an unbounded domain, the definition is analogous, but now the Green's function must also satisfy conditions imposed at ∞.

A consequence of Definition 1.1.5 is that the solution of $\mathcal{L}v = -g$ in Ω and $Bv = 0$ on $\partial\Omega$, is given by

$$v(x) = \int_{supp(g)} G(x,y)g(y)\,\mathrm{d}y, \text{ for all } x \in \Omega. \tag{1.7}$$

We are especially interested in the following boundary value problem and its associated Green's function.

Definition 1.1.6. ***(Half-Plane Problem)*** *Denote by L the line $x_2 = L_2$ and let the half-plane $\Lambda := (-\infty, \infty) \times (-\infty, L_2)$, as in Figure 1-2. Let the wavespeed $c(x)$ be a spatially varying function on Λ such that $(1-c(x))$ has compact support and let $f(x)$ have compact support on Λ. Let $u(x)$ be the solution of the Helmholtz equation*

$$\Delta_x u(x) + k^2(x)u(x) = -f(x), \quad x \in \Lambda, \tag{1.8}$$

with $k(x) := \omega/c(x)$*, with*

$$u = 0, \quad x \in L,$$

with the Sommerfeld radiation condition (SRC)

$$\frac{x}{\|x\|} \cdot \nabla u(x) - \mathrm{i}k(x)u(x) = o\left(\frac{1}{\|x\|}\right) \text{ as } \|x\| \to \infty, \tag{1.9}$$

(observe that k *is constant outside a compact set, so the SRC is well defined).*

Definition 1.1.7. *(*G*) The Green's function* G *(as in Definition 1.1.5) of the half-plane problem in Definition 1.1.6, when* k *is spatially constant, is*

$$G(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{i}}{4} H_0(k\|\mathbf{x} - \mathbf{y}\|) - \frac{\mathrm{i}}{4} H_0(k\|\mathbf{x} - M(\mathbf{y})\|), \tag{1.10}$$

where $M(\mathbf{y})$ *is the reflection of the point* $\mathbf{y}$ *in the line* L *illustrated in Figure 1-2, and* H_0 *is the Hankel function of the first kind.*

For the proof that (1.10) is indeed the Green's function for the half-plane problem, see Definitions 2.2.18 and 2.2.19 and Proposition 2.2.20, where we find Green's functions for a series of half-plane problems.

In this thesis we prove theoretical results for certain Green's functions of Helmholtz problems for spatially constant, but possibly large, wavenumber k. Note that when k is variable there is in general no explicit expression for the fundamental solution, or Green's function. However, the numerical methods motivated by the theory are effective when applied to the more complicated scattering BVPs mentioned above as well, including those with variable $c(x)$ and resulting heterogeneous (or spatially-varying) wavenumber $k(x)$ (see, for example, [35, §3.1-2]).

1.1.3 Motivating Applications

Study of Helmholtz problems is an area of active research with many applications to wave-based phenomena: acoustics [22], elasticity [41], electromagnetics [22] and geophysics [85].

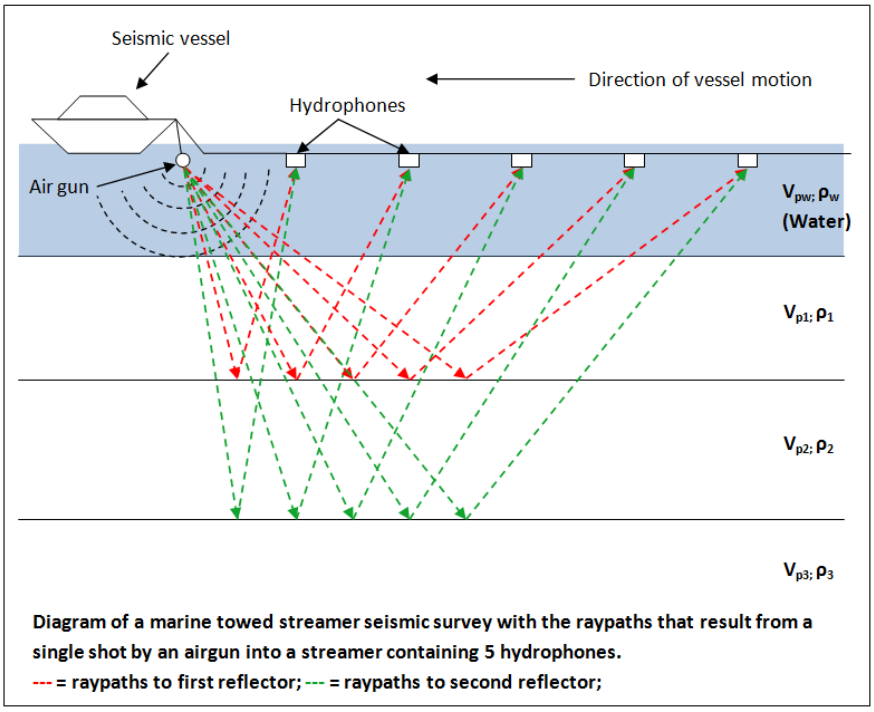


Figure 1-3: Creative commons image [87]. Diagram of a marine seismic survey. The layers of subsurface materials have different velocities v_{pi} and densities ρ_i. At boundaries between the layers the waves are refracted and reflected and these 'echoes' are observed by the hydrophones.

1.1.3.1 Seismic Inversion

This project was sponsored by Schlumberger Cambridge Research because of their interest in seismic inversion.

Seismic inversion aims to find the distribution of parameters describing the physical properties of the earth's subsurface. In the case of the model considered here, the parameter in question will be the wavespeed, as a function of position. To perform seismic inversion, the location must first be surveyed: seismic wave sources (mechanical vibrations on land or an air-gun in the water) are used to generate vibrations/waves in the subsurface area of interest. The pressure echoes emanating from reflections at material boundaries are observed using geophones on land or hydrophones on water. (For an example of the latter see Figure 1-3.) One use of the parameter distribution images created through this process is to see if there is oil at a site and if it can be extracted.

The data sets consisting of the echoes from each source at many frequencies must be interpreted to extract the subsurface parameter information. There are various methods for doing this interpretation, two large categories being tomographic methods and full waveform inversion. Tomographic methods use travel times of rays to determine an approximate velocity model, based on tracing ray paths between source and receiver, assuming the rays go through reflections/refractions based upon possible subsurface parameter distributions, see for example [96, p1750]. However, we focus on full waveform inversion (FWI) as it has the most relevance to our later work. FWI is a non-linear, least-squares minimisation problem, formalised by Lailly [78] and Tarantola [107]. During FWI, the forward problem (of obtaining the scattering pattern of pressure echoes from the sources, given a particular subsurface parameter distribution) is solved many times and the distance between the predicted pressure echoes and the observed pressure echoes is minimised.

The forward problem in FWI can be formulated in the time domain or the frequency domain. We focus on a frequency-domain formulation of FWI, leading to our interest in solving Helmholtz problems. Frequency-domain formulations of FWI can be found in, for example, [47, 95]. A succinct description of the FWI minimisation problem in the frequency domain is given in [85, p155-6]:

$$\min_{m\in\mathcal{M}} A(m) = \min_{m\in\mathcal{M}} \frac{1}{2}\sum_{i=1}^{N_s}\sum_{j=1}^{N_\omega} \|R_i u_i(m,\omega_j) - d_i(\omega_j)\|^2,$$

where

- $A(m)$ is the function to be minimised.
- $\mathcal{M}$ is the space of possible subsurface parameters. For this thesis this is the wavespeed $c(x)$, but in practice it can include more parameters, for example the rock density.
- N_s is the number of sources used at that geographic location.
- N_ω is the number of wave frequencies from the sources.
- $u_i(m,\omega_j)$ is the solution of the forward problem with corresponding parameter $m(x)$, source f_i and frequency ω_j.

- R_i is the mapping of the solution u_i to match the same location as d_i.
- $d_i(\omega_j)$ is the data set of observations corresponding to source f_i and the frequency ω_j.

As we mentioned above, the forward problem is where a wavefield solution/scattering pattern is obtained for a particular source and set of subsurface parameters. A succinct notation for a frequency-domain forward problem is given in [85, p155], as follows:

$$A(m(x), \omega)u(x, \omega) = f(x, \omega) \tag{1.11}$$

where $m \in \mathcal{M}$, x, ω, u and f are as above and $A(m(x), \omega)$ is the partial differential operator for a frequency-domain wave problem, which could be the simplest problem with the Helmholtz operator, or a much more complicated variant; we discuss the possibilities for this operator shortly.

Solving the forward problem must be done many times, for three reasons. Firstly there will be many different sources or forcing terms used to generated any image. Secondly there will be multiple frequencies to solve for. Finally as the parameter distribution is determined by gradually improving an initial guess of the distribution, until the predicted pressure echoes match the observations as closely as possible, the forward problems for all the different sources and frequencies must be solved many times. Hence an efficient way to calculate the solution to the forward problem is of wide interest.

As alluded to in the description of (1.11), the formulation of the forward problem can take several different forms. For example a very general formulation is the poro-elastic wave model developed by Biot [9], which takes into account of various properties of the subsurface rocks, like elasticity, porosity and viscosity. However, a balance must be struck between the complexity of the model and what can effectively be modelled on a useful time-scale with the available computing resources. Generally a much simpler model is used.

When we consider only linear, isotropic elasticity, we have to solve the elastodynamic wave equation:

$$\rho\frac{\partial^2 u}{\partial t^2} - \mathrm{div}C\epsilon(u) \ = \rho\frac{\partial^2 u}{\partial t^2} - \mathrm{div}(\lambda \mathrm{div}(u)\mathbb{I} + 2\mu\epsilon(u)) = \ f, \tag{1.12}$$

see, for example, [41, p19-20], where

- $u(x,t)$ is the displacement: after time t, the position of the particle at x is given by $x'(x,t) = x + u(x,t)$,
- $\epsilon(u(x,t))$ is the strain tensor,
- C is the stiffness tensor, which depends on Lamé elasticity parameters λ, μ, and
- $\rho(x)$ is the density.

(Note that isotropic elasticity means that it has the same elastic properties in every direction.)

The elastodynamic equation models both pressure (longitudinal) and shear (transverse) waves (P- and S-waves respectively). The elastodynamic wave equation does not model memory effects, or attenuation of the waves (i.e., the waves losing energy and amplitude as they pass through the rocks), viscosity or porosity [41, p17]. (Attenuation can by modelled in the frequency domain by using a complex frequency in the equation [29], a potential application of our later work where we consider adding a complex part to the wavenumber.)

The acoustic approximation is a further approximation that can be made to the elastodynamic equation. In the acoustic approximation only P-waves are considered, so that effectively, the shear velocity is being set to 0 [41, p22]. The acoustic approximation is just as useful as the elastodynamic equation in certain situations. Since the P-waves travel faster and arrive at the receivers first, the acoustic approximation does not affect methods that only involve first-time arrivals [41, p21]. The acoustic approximation is beneficial in that it is cheaper to compute and simpler to analyse and requires a smaller grid upon which to do computations. The S-wave wavelength *"is at least 1.5 times smaller than the P-wave wavelength and therefore the elastic model requires a finer mesh"* [41, p21 and references therein]. Limitations of the acoustic approximation to the elastic wave equation are discussed in, for example, [12].

We describe the acoustic approximation, where shear waves are neglected, by following [41, p22-23]. Let C be the stiffness tensor and $p = \kappa \text{div}(u)$ be the pressure, where κ is the bulk modulus. In, for example, [41, p22-23], it is shown

that

$$C\epsilon(u) = p\mathbb{I}, \tag{1.13}$$

and thus

$$\mathrm{div}C\epsilon(u) = \nabla p. \tag{1.14}$$

Then (1.12) simplifies to

$$\rho\frac{\partial^2 u}{\partial t^2} - \nabla p = f. \tag{1.15}$$

We divide (1.15) through by the density ρ and also take the divergence to obtain:

$$\mathrm{div}\frac{\partial^2 u}{\partial t^2} - \mathrm{div}\left(\frac{1}{\rho}\nabla p\right) = \mathrm{div}\left(\frac{1}{\rho}f\right). \tag{1.16}$$

By the definition of the pressure ($p = \kappa\mathrm{div}(u)$ above) and interchanging the order of derivatives we obtain

$$\mathrm{div}\frac{\partial^2 u}{\partial t^2} = \frac{1}{\kappa}\frac{\partial^2 p}{\partial t^2}. \tag{1.17}$$

Finally we substitute (1.17) into (1.16) to obtain

$$\frac{1}{\kappa}\frac{\partial^2 p}{\partial t^2} - \mathrm{div}\left(\frac{1}{\rho}\nabla p\right) = \mathrm{div}\left(\frac{1}{\rho}f\right). \tag{1.18}$$

Equation (1.18) is the time domain acoustic model. To obtain the frequency-domain formulation we take the Fourier transform of (1.18) in time.

Let the Fourier transform of a function $\psi(x,t)$, that satisfies certain regularity assumptions, with respect to time be

$$\mathcal{F}(\psi)(x,\omega) = \int_{-\infty}^{+\infty} \psi(x,t)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}t,$$

Note that

$$\mathcal{F}\left(\frac{\partial\psi}{\partial t}\right) = -\mathrm{i}\omega\mathcal{F}(\psi).$$

Denoting the Fourier transforms of p and f by p_ω and f_ω, by (1.18) we have

$$-\frac{\omega^2}{\kappa^2}p_\omega - \mathrm{div}\left(\frac{1}{\rho}\nabla p_\omega\right) \;=\; \mathrm{div}\left(\frac{1}{\rho}f_\omega\right), \tag{1.19}$$

where ω is the Fourier frequency. Solving (1.19) for each ω and doing reverse Fourier transform then yields a time domain solution to (1.18).

In practice the forward problem (1.18) is often solved in the time domain, but there is continued interest in frequency-domain solvers, for example [85].

A final simplification that can be made to (1.19) is to assume constant density. We then obtain a variant of the Helmholtz equation as follows

$$-\frac{1}{c_p^2}p_\omega - \Delta p_\omega = \mathrm{div} f = F.$$

Hence the Helmholtz equation is arguably the simplest approximation to the wave equation that it is useful to consider and experiment with in the context of FWI.

1.2 Discretisation Methods

1.2.1 Methods and Costs

To discretise Helmholtz problems there are two main categories of methods to chose from: volume discretisation methods and boundary integral equation methods.

Volume methods are characterised by the fact that the problem is formulated on a mesh on the domain that the PDE problem is posed on. Some common examples are finite difference, finite element and finite volume methods. In this thesis we use volume methods for all of our numerical experiments. Unlike boundary integral methods, volume methods do not require prior knowledge of the Green's function, which is important for our application of interest as the Green's function is not generally known in practice for inhomogeneous problems.

Finite difference methods use difference equations to approximate the derivatives. In practice high-order methods are often used (see for example [103]), though for simplicity our demonstrative numerical tests use low-order methods, see description in §2.1.2.

Finite element methods subdivide the domain into pieces called finite elements, and discretise a variational/weak form of the PDE upon the mesh of finite elements to create a linear system of equations. We create such a variational formulation of the Helmholtz equation in §1.4.2 and details of the exact variant of the method that we use in numerical experiments are given in §2.1.3.

For further details about finite element discretisations of Helmholtz problems we refer the reader to [69].

The number of degrees of freedom in volume methods is generally $\mathcal{O}(n^d)$, where d is the dimension of the domain in which the PDE problem is posed and n is the maximum number of degrees of freedom along the length of any dimension. Volume methods generally yield sparse matrices (when the method considers only local interactions) of a size $\mathcal{O}(n^d) \times \mathcal{O}(n^d)$.

In practice, n (the maximum number of degrees of freedom along the length of any dimension) has to be large for Helmholtz problems. To see this, we first note that the important factor about the size of the domain that Helmholtz problems are posed on, is the characteristic length kL, i.e. the maximum size in any dimension of the domain L, multiplied by the wavenumber k. This characteristic length gives a measure of the number of waves within the domain or how 'oscillatory' the problem is. The characteristic length can be large either due to k being large, or L being large. In this thesis, for simplicity we assume that k is the large parameter and L is $\mathcal{O}(1)$. Solutions to Helmholtz problems are then oscillatory with wavelength $\mathcal{O}(1/k)$. For these solutions to be resolved one needs a fixed number of points of per wavelength, i.e., $h = \mathcal{O}(1/k)$, where h is the mesh size. This small mesh size results in linear systems for a volume method with a matrix of size $\mathcal{O}(k^d) \times \mathcal{O}(k^d)$ in dimension d (note that $h = \mathcal{O}(1/n)$). In fact, when low-order finite difference or finite element methods are used, accuracy is not maintained as k increases for a fixed number of points per wavelength due to a phenomenon known as the pollution effect (see §1.4.2). Therefore linear systems with $h = \mathcal{O}(1/k^\alpha)$, with $\alpha > 1$, are required to maintain accuracy, and these linear systems are therefore even larger in size: $\mathcal{O}(k^{d\alpha}) \times \mathcal{O}(k^{d\alpha})$.

We briefly mention some details about the second category of methods, integral equation methods, as they are indirectly of interest in this thesis. When the fundamental solution is known, the problem can be reformulated as an integral equation on the boundary of the domain; we refer the reader to details in, for example, [22, §3 especially p39-48]. When the Galerkin method is used to solve the integral equation, the resulting method is known as the Boundary Element Method (BEM), see for example [76]. A recent overview describing boundary integral methods for Helmholtz problems may be found in [13]. The boundary integral equation methods can deal especially well with radiation conditions as

they are included in the boundary integral formulation. However, we note that, due to the need for prior knowledge of the fundamental solution, the methods can only be applied to homogenous, or piecewise homogeneous media $c(x)$. Whilst there are many applications for which this assumption holds true (see for example many applications in [16, §4]), there are limitations to the methods' usefulness, for example, when it comes to seismic inversion problems as described in §1.1.3.1, where the goal is to model the wavespeed distribution $c(x)$.

Since discretisation only takes place on the boundary, the number of degrees of freedom in boundary integral equation methods is $\mathcal{O}(n^{d-1})$, where d is the dimension of the domain in which the PDE is posed and n is the maximum number of degrees of freedom along the length of any dimension.

To fully explain the discretisations of the model problems we consider in this thesis, we need to consider two further things: 1) in §1.3 we consider how to discretise/approximate the Sommerfeld Radiation Condition (1.6) and 2) in §1.4.2 we consider the pollution effect which for low-order methods means that a very fine grid is needed to successfully compute a solution.

1.3 Approximating Sommerfeld Radiation Condition

In this thesis we are interested in solving scattering Helmholtz problems, posed with the Sommerfeld Radiation Condition. When discretisation is done with a volume method like the finite difference method or the finite element method, the domain of the problem must be truncated to some finite area of interest. Thus the Sommerfeld Radiation Condition (1.6) must be approximated. A whole family of methods have been developed to do this approximation; we discuss just two of these.

1.3.1 Absorbing Boundary Conditions

Engquist and Majda [32] created some absorbing boundary conditions that are Padé approximations to localisations of the pseudo-differential operator that provides the exact boundary condition. The impedance boundary condition is the

zeroth-order Padé approximation and takes the form:

$$\frac{\partial u}{\partial n} - \mathrm{i}\gamma u = 0, \quad \text{on } \partial\Omega,$$

for some constant γ, usually chosen as k [32], [69, §3.3.2].

1.3.2 Perfectly Matched Layer

A much more sophisticated method is to use a Perfectly Matched Layer (PML). Introduced by Berenger [8], the idea of the PML is nicely summarised by Ihlenburg [69, §3.3.4]: *"the idea is to introduce an exterior layer at the artificial boundary in such a way that all plane waves are totally absorbed* [by being forced to decay exponentially within this layer]. *This means that no reflection occurs... and the transmitted wave vanishes at infinity, whence the name perfectly matched layer (PML) method. In practice, the computation is truncated at some finite distance within the layer. But the resulting artificial reflections are small, due to the exponential decay."* Or in other words, within the PML region we have exponential decay of incident waves [20] [69, §3.3.4], mimicking the Sommerfeld radiation condition and aiming to ensure that there are no reflections.

In this thesis we use PMLs for all our numerical experiments as they are simple to implement and effective in the cases we consider, see details in §2.1.1.

1.4 Finite Element Method

1.4.1 Solution of Interior Impedance Problem

In this thesis, we are concerned with discretisations of Helmholtz problems created with low-order finite element methods or finite difference methods. In this section we briefly describe the finite element method.

Definition 1.4.1. *(**Function spaces** L^2**,** H^1*) The space $L^2(\Omega)$ is the space of square integrable functions $f : \Omega \to \mathbb{C}$, i.e., functions f such that*

$$\|f\|_2^2 := \int_\Omega |f|^2 \,\mathrm{d}x,$$

exists and is bounded. (Equipping this with the norm $\|\cdot\|_2$ makes it a Hilbert space.)

The space $H^1(\Omega)$ is the Sobolev space

$$H^1(\Omega) := W^{1,2}(\Omega) := \{u \in L^2(\Omega) \,\Big|\, D^\alpha u \in L^2(\Omega), \text{ for all } \alpha \in \mathbb{N}_0^d : |\alpha| \leq 1\},$$

where derivatives

$$D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}},$$

are defined in a distributional sense.

Since the details of Sobolev spaces are only of incidental interest in this thesis, we refer the reader to, for example, [79] for full details.

Definition 1.4.2. ***(Interior Impedance Problem (IIP))*** *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded, Lipschitz, open set and let $\Gamma = \partial\Omega$. Given $f \in L^2(\Omega)$ $g \in L^2(\Gamma)$ and $k > 0$, find $u \in H^1(\Omega)$, the solution of the Helmholtz equation*

$$\Delta_x u(x) + k^2 u(x) = -f(x), \quad x \in \Omega, \tag{1.20}$$

with the impedance boundary condition

$$\partial_n u(x) - \mathrm{i}k u(x) = g(x), \qquad x \in \Gamma, \tag{1.21}$$

where ∂_n denotes the normal derivative operator.

Definition 1.4.3. ***(Standard Variational Formulation of the IIP)*** *Given $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$ and $k > 0$, find $u \in H^1(\Omega)$ such that*

$$a(u, v) = F(v), \quad \text{for all } v \in H^1(\Omega),$$

where

$$a(u, v) := \int_\Omega \nabla u(x) \cdot \overline{\nabla v(x)} - k^2 u(x)\overline{v(x)}\,\mathrm{d}x - \mathrm{i}k \int_\Gamma u(x)\overline{v(x)}\,\mathrm{d}x \tag{1.22}$$

and

$$F(v) := \int_\Omega f(x)\overline{v(x)}\,\mathrm{d}x + \int_\Gamma g(x)\overline{v(x)}\,\mathrm{d}x.$$

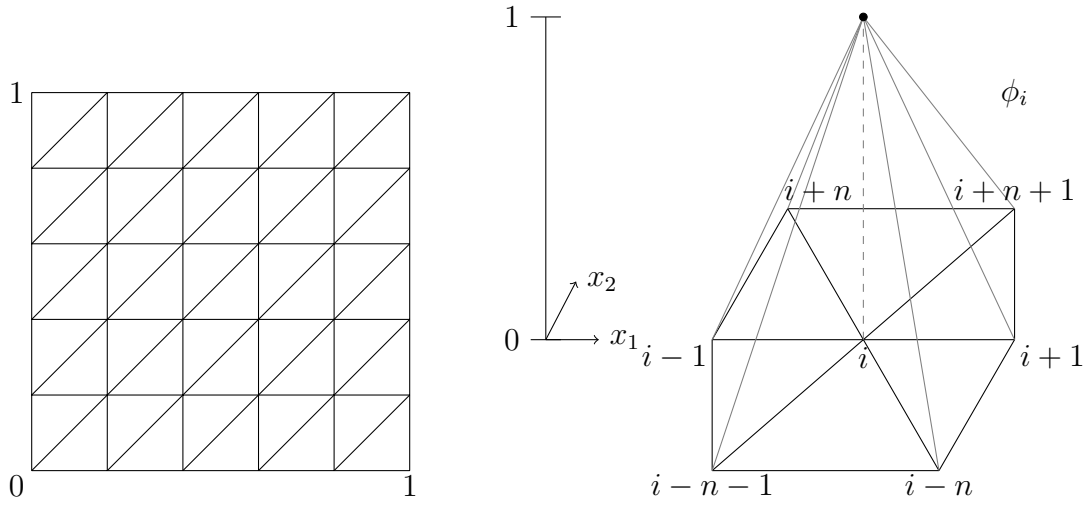


Figure 1-4: Left: Example of a grid for finite elements for $\Omega = [0,1]^2$. Right: if $p = 1$, a piecewise polynomial ϕ_i is known as a hat function. The diagram shows the hat function for grid node x_i. Note that for compactness we have written, for example, i or $i-n$ for nodes x_i or x_{i-n} respectively.

The standard variational formulation is derived by multiplying the Helmholtz problem of Definition 1.4.2 by a test function $\bar{v} \in H^1(\Omega)$ and using Green's theorem to integrate by parts.

Definition 1.4.4. ***(Galerkin equations)*** *Given a finite dimensional subspace $T_N \subset H^1(\Omega)$, find $u_N \in T_N$ such that $a(u_N, v_N) = F(v_N)$, for all $v_N \in T_N$.*

In the finite element method (FEM) there are several possible choices for T_N. Recall that the FEM is formed by creating a meshed partition of the domain of the PDE problem with covering elements K_i, $i \in \{1, \ldots, N_K\}$ and with nodes x_i, $i \in \{1, \ldots, N\}$. We define the maximum distance between any two nodes of the mesh to be h. For a regular mesh, $N \sim h^{-d}$ is the total number of degrees of freedom in the subspace. Then for any $p \geq 1$, we define T_N to be the space of all continuous functions which reduce to polynomials of degree p on each element of the mesh. Let a basis for T_N be denoted $\{\phi_i : i \in \{1, \ldots, N\}\}$. Now we can reformulate the Galerkin equations from Definition 1.4.4 as the following linear system of equations.

Definition 1.4.5. ***(Matrix A from FEM)*** *Let the linear system that arises from the FEM discretisation of the Interior Impedance Problem (Definition 1.4.2)*

be denoted by

$$A\boldsymbol{u} = \boldsymbol{f}, \tag{1.23}$$

where $A \in \mathbb{C}^{N\times N}$, $\boldsymbol{u} \in \mathbb{C}^{N\times 1}$ *and* $\boldsymbol{f} \in \mathbb{C}^{N\times 1}$ *are such that*

$$A = S - k^2 Q - \mathrm{i}kP,$$

where the matrices and vectors are defined as follows: the stiffness matrix is $S_{i,j} := \int_\Omega \nabla\phi_i.\nabla\phi_j \,\mathrm{d}x$, *the domain mass matrix is* $Q_{i,j} := \int_\Omega \phi_i\phi_j \,\mathrm{d}x$, *the boundary mass matrix is* $P_{i,j} := \int_\Gamma \phi_i\phi_m \,\mathrm{d}s$, *the solution vector is* $\boldsymbol{u}_i := \int_\Omega u_N\phi_i \,\mathrm{d}x$ *and the source vector is* $\boldsymbol{f}_i := \int_\Omega f\phi_i \,\mathrm{d}x + \int_\Gamma g\phi_i \,\mathrm{d}s$.

Remark 1.4.6. *The matrix* A *is symmetric but not Hermitian, because of the* i *in front of the matrix* P. *This* i *appears because of the impedance boundary condition* (1.21). *Recall from §1.3.1 that the impedance boundary condition is the simplest Absorbing Boundary Condition that approximates to the Sommerfeld Radiation condition. Therefore the non-self-adjointness of the matrix* A *is consistent with the fact that the Helmholtz problem with the Sommerfeld Radiation condition is not self-adjoint at the PDE-level, see Remark 1.1.4. This will hold true for other discretisations of Helmholtz problems with the Sommerfeld Radiation condition as well.*

1.4.2 Pollution Effect

1.4.2.1 Definition of the Pollution Effect

Discretisations of Helmholtz problems from low-order methods, like (1.23) from the FEM with a small value of polynomial degree p, suffer from a phenomenon known as the pollution effect. In order to explain this phenomenon, we consider the pollution effect in the case of the example of the h-FEM discretisation of the interior impedence problem. We begin our description of the pollution effect with some necessary details about the analysis of the FEM for the Helmholtz problems in terms of h and k, following, for example, [28, §2].

Numerical analysis of the FEM shows that when the method is well formulated, by decreasing h, increasing p, or both (the basis of the h-FEM, p-FEM or hp-FEM respectively), the solution vector $\boldsymbol{u}$ will converge to the solution u of the

PDE problem. However, for Helmholtz problems, the difficulty of the problem is largely determined by the size of the wavenumber k and so it is necessary to consider the convergence analysis for FEM with h and p in relation to k. In this thesis, we perform numerical experiments only with h-FEM with $p = 1$ and so our description of the analysis for the pollution effect are focused on this case.

Definition 1.4.7. ***(Helmholtz norm*** $\|\cdot\|_{H^1_k(\Omega)}$***)*** *Let the Helmholtz norm be defined as*

$$\|v\|^2_{H^1_k(\Omega)} := \|\nabla v\|^2_{L^2(\Omega)} + k^2\|v\|^2_{L^2(\Omega)}.$$

The Helmholtz norm $\|\cdot\|_{H^1_k(\Omega)}$ is the natural one for Helmholtz problems as the first order derivative of any of the fundamental solutions (see Definition 1.1.1.1) features a power of k. It is therefore to be expected that $\|\nabla u\|_{L^2(\Omega)} \sim k\|u\|_{L^2(\Omega)}$, when u is the solution of the Helmholtz equation and so the k^2 factor in $\|\cdot\|_{H^1_k(\Omega)}$ should make the two terms in the norm the same order of magnitude. (Note that we use $\sim$ in the following sense: if $a, b > 0$ we write $a \lesssim b$ if $a \leq Cb$ for some $C > 0$ that is independent of all quantities of interest, then $a \sim b$ if $a \lesssim b$ and $b \lesssim a$.)

The first definition of the pollution effect relates to the quasi-optimality of the method.

Definition 1.4.8. ***(Quasi-optimality)*** *The h-FEM is quasi-optimal if the solution u_N of the FEM satisfies*

$$\|u - u_N\|_{H^1_k(\Omega)} \leq C_{qo} \min_{v_N \in H_N} \|u - v_N\|_{H^1_k(\Omega)} \tag{1.24}$$

for some constant C_{qo} that may depend on h and k.

The FEM is said to suffer from the pollution effect if $hk \sim 1$ (a grid spacing with fixed number of points per wavelength) is not sufficient to guarantee that (1.24) holds with C_{qo} independent of h and k. For various situations, it has been proved, or numerical experiments have been conducted to show, that $h \sim k^{-2}$ is sufficient for the h-FEM to be quasi-optimal with C_{qo} independent of h and k, see references in Remark 1.4.10.

The second definition of the pollution effect relates to the relative error of the method.

Definition 1.4.9. ***(Relative error)*** *The relative error of the h-FEM is as follows:*

$$\textit{Relative error} := \frac{\|u - u_N\|_{H^1_k(\Omega)}}{\|u\|_{H^1_k(\Omega)}} \tag{1.25}$$

where u_N is the FEM solution.

The FEM is said to suffer from the pollution effect if $hk \sim 1$ (a grid spacing with a fixed number of points per wavelength) is not sufficient to ensure that the relative error (1.25) is bounded as $k \to \infty$. For various situations, it has been proved, or numerical experiments have been conducted to show, that $h \sim k^{-3/2}$ is sufficient to keep relative error bounded as $k \to \infty$, see references in Remark 1.4.10.

Remark 1.4.10. ***More information about the two definitions of the pollution effect***

For more complete summaries of results about the pollution effect with respect to quasi-optimality, we refer the reader to [28, p4] and [52, p182-183]. Here we note that the pollution effect with respect to the quasi-optimality in 1D was examined by Ihlenburg and Babuška, who proved that when hk^2 is sufficiently small, for $p = 1$, the h-FEM, for the problem with impedance condition at one end and Dirichlet condition at the other, is quasi-optimal with C_{qo} independent of h and k, see [70, Theorem 3], [69, Theorems 4.9 and 4.13] and their numerical experiments suggest that the requirement $h \sim k^{-2}$ is sharp [70, Figures 7 and 8], [69, Figure 4.11]. The situation in 2D and 3D is less well understood. In 2D and 3D, when hk^2 is sufficiently small, Melenk proved (under regularity conditions on the solution and other conditions on the solution involving the data of the problem f and g and the diameter of Ω) that the h-FEM is quasi-optimal with C_{qo} independent of h and k, see [83, Proposition 8.2.7].

For a more complete summary of the results about the pollution effect with respect to the relative error, we refer the reader to [28, p3-4]. Here we note that the pollution effect with respect to the relative error in 1D was examined by Bayliss, Goldstein and Turkel [5] and then more completely by Ihlenburg and Babuška [72]. When $hk^{3/2} \sim 1$, when $p = 1$, a H^1-conforming piecewise polynomial subspace FEM for a PDE problem with $u \in H^2$ has strictly decreasing relative error for all $k \geq k_0$, for some $k_0 > 0$, [70, Equation 3.25], [69, Equation 4.5.15] and their

numerical experiments suggest that the requirement $h \sim k^{-3/2}$ is sharp [72, Figure 11], [69, Figure 4.13].

When $d = 2$ or 3, far less is proved about the pollution effect with respect to the relative error. For a restricted set of problems, Wu obtained a bound for $\|u - u_N\|_{H^1_k(\Omega)}$ in terms the data f and g, for more details see [113]. For another, less restricted set of problems Melenk and Sauter showed that a similar bound (to that in [113]) holds [84, Equation (5.14b)], when hk^2 is sufficiently small.

For more details about the pollution effect, we refer the reader to [71] (for hp-FEM discretisations) and [52, §1.2] and references therein (for BEM discretisations and a comparison of the pollution effect for BEM and FEM discretisations).

1.4.2.2 Demonstration of Pollution Effect

We demonstrate the pollution effect with a numerical experiment that investigates how the pointwise relative error of the finite element implementation of 2D Helmholtz problems varies with k and h.

We solve the interior impedence problem in Definition 1.4.2 on $\Omega = [0,1]^2$, where g is such that the solution the plane wave $u(x) = \exp(\mathrm{i}kx.a)$ where $a := (1/\sqrt{2}, 1/\sqrt{2})^T$. We solve it using the linear FEM, with the standard formulation in Definition 1.4.3. We use triangular, piecewise-linear finite elements and hat functions as in Figure 1-4. We compute the integrals using the composite centroid rule on triangles.

Since the exact solution (the plane wave) is known, we are able to use the exact solution and the numerical solution u_N to calculate the pointwise relative error at $(1/2, 1/2)$:

$$\begin{aligned}\text{Pointwise Relative Error } &= \frac{u(1/2,1/2) - u_N(1/2,1/2)}{|u(1/2,1/2)|} \\ &= \exp\left(\frac{\mathrm{i}k}{\sqrt{2}}\right) - u_N(1/2,1/2).\end{aligned}$$

For each k we decrease h (or equivalently increase N) until the pointwise relative error is within a fixed tolerance ϵ. Figure 1-5 is a plot of $\log h$ verses $\log k$ for three different tolerances $\epsilon = \{0.1, 0.01, 0.001\}$. We fit linear trend-lines and find that the relationship between k and h needed to maintain a fixed pointwise relative error is approximately $h \sim k^{-3/2}$. Our experiment shows good agreement with

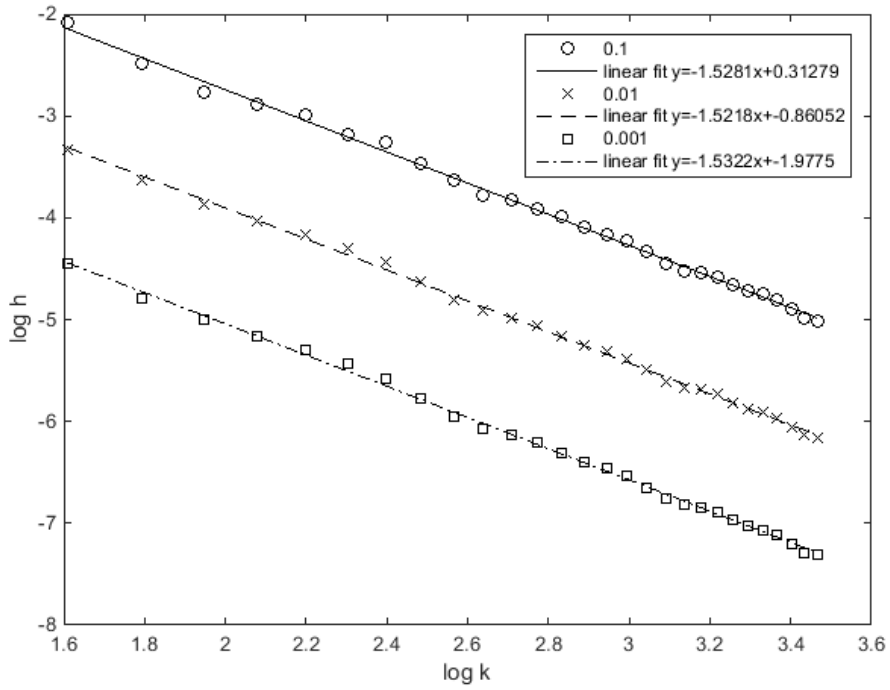


Figure 1-5: Pointwise fixed relative error for h and k, for values of ε given in the legend. The trendlines show that $h \sim k^{-3/2}$ is need to maintain the fixed pointwise relative error.

similar experiments in [113, §7].

1.4.2.3 Consequences of the Pollution Effect

The evidence in §1.4.2 suggests that in order to counteract the pollution effect a grid spacing of $h \sim k^{-\mu}$, for somewhere in the range $1 \leq \mu \leq 2$ may be required, depending on the order of the method used and the particular problem being solved. Later in this thesis, when we need to consider the range of values that h may take in terms of k, we use this range of possible values. We also conduct numerical experiments, for these we use $h \sim k^{-1}$ and $h \sim k^{-3/2}$.

For now, we note that the small mesh sizes required to counteract the pollution effect result in discretisations of Helmholtz problems having linear systems like (1.23) to solve that are very large. In fact, the discretisation matrices like A in (1.23), must be of the size $\mathcal{O}(k^{\mu d}) \times \mathcal{O}(k^{\mu d})$ for $1 \leq \mu \leq 2$ for dimension d. Hence it is of particular importance and interest to find efficient solution methods

for linear systems arising from Helmholtz problems and we give an overview of methods of interest to this thesis in the next few sections.

1.5 Solution of the Linear System (1.23)

1.5.1 Direct Methods

For a linear system with N unknowns, the cost to invert the linear system via Gaussian elimination is $\mathcal{O}(N^3)$. However, we recall that the linear systems that arise from the discretisation of Helmholtz problems are extremely large, due to the need to resolve the waves and counteract the pollution effect for low-order methods (see §1.2 and §1.4.2), and so a cost of $\mathcal{O}(N^3)$ is prohibitively high for discretisations of Helmholtz problems and we want a cost as close to $\mathcal{O}(N)$ as possible.

There are many direct methods for solving linear systems arising from integral-equation discretisations of PDE problems (for example, Laplace problems), that have costs much less than $\mathcal{O}(N^3)$, see for example [63,77,81]. Most of these integral equation direct methods assume certain properties of the matrix, for example saying that off-diagonal or 'admissible' matrix blocks are 'rank deficient' or 'compressible', meaning that they readily admit low-rank approximations. However oscillatory Helmholtz kernels do not always have the equivalent of this property (i.e., the kernels do not readily admit low-rank separable expansions, see discussion in §1.8), especially for high-wavenumber and complicated geometries. Therefore the associated matrices do not have this property of readily admitting low-rank approximations and so more thought must be put in to applying these methods to Helmholtz problems than to Laplace problems. For example, in [82], Rokhlin and Martinsson adapt their method from [81] for non-oscillatory kernels to work for a Helmholtz problem in the particular situation of elongated scatterers in [82]. In this particular scenario, the Helmholtz kernel and hence the off-diagonal matrix blocks in the linear system, can be shown to be low-rank, even for high frequencies.

There are also direct methods for solving volume discretisations of Helmholtz problems that also exploit low-rank properties (see §1.8), for example those by Gillman, Martinsson et al. [48, 49]. These methods perform well for low-to-

medium frequency problems, but high-frequency problems are still challenging, as in many cases the cost to construct or apply the methods is $\mathcal{O}(N^\alpha)$ for $\alpha > 1$.

Direct methods have the following advantage when there are multiple right-hand sides: the direct method to solve the system only needs be constructed once and can then be used for each right-hand side. Therefore they are more likely to be competitive in situations with multiple right-hand sides.

1.5.2 Iterative Methods

In this thesis we mostly focus on solving the Helmholtz problem using iterative methods. An iterative method to solve the $N \times N$ linear system

$$A\boldsymbol{x} = \boldsymbol{b},$$

is a method that generates a series of approximate solutions $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, that converge to the solution $\boldsymbol{x}$. Each $\boldsymbol{x}_i$ is obtained from the previous iterate(s) by some multiplications with A.

There are many different iterative methods that are appropriate in different situations, depending, for example, upon the size of the linear system considered and the properties of the system matrix A. A common choice of iterative methods for solving systems with large, sparse matrices arising from Helmholtz problems is to use iterative methods based on Krylov subspaces, see Definition 1.5.1, with some form of preconditioner, see §1.6.

Definition 1.5.1. ***(Krylov Subspace** K_l**)*** *Let $A \in \mathbb{C}^{N\times N}$ and $\boldsymbol{v} \in \mathbb{C}^N$. Then for $l \in \mathbb{N}$, $l < N$, the lth Krylov subspace is*

$$K_l(A, \boldsymbol{v}) = \text{span}\{\boldsymbol{v}, A\boldsymbol{v}, A^2\boldsymbol{v}, \ldots, A^{l-1}\boldsymbol{v}\}, \tag{1.26}$$

where $\boldsymbol{v} := \boldsymbol{r}_0 / \|\boldsymbol{r}_0\|$, where $\boldsymbol{r}_0 := \boldsymbol{b} - A\boldsymbol{x}_0$, for an initial guess $\boldsymbol{x}_0$.

There are two classes of Krylov subspace methods of interest in this thesis. Of most direct interest is a class of methods that includes the Generalized Minimal Residual method (GMRES) [99, 100]. (Note that GMRES is a generalisation of the Minimal Residual Method (MINRES) [94]. However, as MINRES requires a Hermitian system matrix, it is GMRES that is relevant to the linear

systems that we consider.) Of less direct interest is a second class of iterative methods, that includes the Conjugate Gradient Method (CG) [67] (which also requires Hermitian matrices), and its generalisation to non-Hermitian matrices the Bi-Conjugate Gradient Method (Bi-CG) [39] and a modified version of Bi-CG, Conjugate Gradient Squared (CGS) [104]. A further generalisation of CGS and the most effective of the well known algorithms in this class is Bi-Conjugate Gradient Stabilised Method Bi-CGSTAB [110]. (Bi-CGSTAB can be thought of as a blend of Bi-CG and GMRES [74, p50], where one iteration of Bi-CG is followed by one iteration of GMRES restarted (see §1.5.2.1 below for restarted GMRES) at each step [74, p50], [4, p27] and [88, p33].) Bi-CG, CGS and Bi-CGSTAB can all suffer breakdown in some situations [74, p47-9] and [111, p137], but as Bi-CGSTAB is the most well established we consider this method here rather than the others.

GMRES is the method of choice for the numerical experiments in this thesis and we give an outline of it below. (For an explanation of the choice of this method, see §1.5.2.2.)

1.5.2.1 GMRES

GMRES finds the lth iterate $\boldsymbol{x}_l$, where $\boldsymbol{x}_l \in \boldsymbol{x}_0 + K_l(A, \boldsymbol{r}_0)$ and $\mathbf{x}_l$ is defined such that $\boldsymbol{b} - A\boldsymbol{x}_l \perp AK_l(A, \boldsymbol{r}_0)$, so that the residual norm over vectors in $\boldsymbol{x}_0 + K_l(A, \boldsymbol{r}_0)$ is minimized [99, p159,p164]. To do calculate $\mathbf{x}_l$, an orthogonal basis for the Krylov subspace $K_l(A, \boldsymbol{r}_0)$ is calculated using the Arnoldi method and the minimisation of the residual is done using this basis. A basic version of the algorithm for GMRES that uses the Modified Gram-Schmidt (MGS) orthogonalisation in the Arnoldi method is as follows.

Algorithm 1.5.2. ***GMRES Algorithm*** *[99, Algorithm 6.9]*

1: Compute $\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}_0$, $\beta = \|\boldsymbol{r}_0\|$ *and* $\boldsymbol{v}_1 = \boldsymbol{r}_0/\beta$

2: for $j = 1, 2, \ldots, l$

3: *Compute* $\boldsymbol{w}_j = A\boldsymbol{v}_j$

4: *for* $i = 1, \ldots, j$ *% Loop contains the Arnoldi method with MGS*

5: $h_{ij} = (\boldsymbol{w}_j, \boldsymbol{v}_i)$

6: $\boldsymbol{w}_j = \boldsymbol{w}_j - h_{ij}\boldsymbol{v}_i$

7: *end*

8: $h_{j+1,j} = \|\boldsymbol{w}_j\|_2$. *If* $h_{j+1,j} = 0$ *set* $l = j$ *and go to 11.*

9: $\boldsymbol{v}_{j+1} = \boldsymbol{w}_j / h_{j+1,j}$

10: *end*

11: *Define the* $(l+1) \times l$ *Hessenberg matrix*

$$\tilde{H}_l = \{h_{ij}\}_{1 \le i \le l+1, 1 \le j \le l}$$

12: *Compute* $\boldsymbol{y}_l$, *the minimiser of* $\|\beta \boldsymbol{e}_1 - \tilde{H}_l \boldsymbol{y}\|_2$ *and* $\boldsymbol{x}_l = \boldsymbol{x}_0 + V_l \boldsymbol{y}_l$.

(Note that we define $\mathbf{e}_l$ to be the lth standard basis vector.)

To gain some additional understanding of the Algorithm 1.5.2, in particular how the each iterate $\mathbf{x}_l$ is calculated to minimise the residual norm over vectors in $\boldsymbol{x}_0 + K_l(A, \boldsymbol{r}_0)$ (see especially Algorithm 1.5.2 Step 12), we follow some theory from [99].

Proposition 1.5.3. *[99, Proposition 6.5] Let* V_l *be the* $N \times l$ *matrix with column vectors* $[\mathbf{v}_1, ..., \mathbf{v}_l]$ *(vectors as in Algorithm 1.5.2), and let* H_l *be* $\widetilde{H}_l$ *with its last row removed.*
Then

$$AV_l = V_{l+1}\widetilde{H}_l$$

and

$$V_l^T A V_l = H_l \tag{1.27}$$

hold.

We can then see that

$$\begin{aligned} \mathbf{b} - A\mathbf{x} &= \mathbf{b} - A(\mathbf{x}_0 + V_l \mathbf{y}) \\ &= \mathbf{r}_0 - AV_l \mathbf{y} \\ &= \beta \mathbf{v}_1 - V_{l+1}\widetilde{H}_l \mathbf{y} \\ &= V_{l+1}(\beta \mathbf{e}_1 - \widetilde{H}_l \mathbf{y}), \end{aligned}$$

which, as V_{l+1} is orthonormal by the construction of the vectors $\mathbf{v}_i$ in Algorithm 1.5.2, implies

$$\|\mathbf{b} - A\mathbf{x}\|_2 = \|\mathbf{b} - A(\mathbf{x}_0 + V_l\mathbf{y})\|_2 = \|\beta\mathbf{e}_1 - \widetilde{H}_l\mathbf{y}\|_2, \tag{1.28}$$

[99, p164-5]. Therefore, $\mathbf{x}_l = \mathbf{x}_0 + V_l\mathbf{y}_l$, where

$$\mathbf{y}_l = \text{argmin}_{\mathbf{y}}\|\beta\mathbf{e}_1 - \widetilde{H}_l\mathbf{y}\|$$

[99, p165]. So $\mathbf{x}_l$ does indeed minimise of the residual norm over vectors in $\boldsymbol{x}_0 + K_l(A, \boldsymbol{r}_0)$.

There are many more advanced versions of the GMRES algorithm, for example using Householder orthogonalization in the Arnoldi method [99, §6.5.2], which is more numerically stable than MGS (sometimes even with MGS, there is 'severe' cancellation of the $\boldsymbol{w}_j$'s in the Arnoldi method) [99, p 156]. (The leading order cost of the Arnoldi method with any of these variants is $\mathcal{O}(l^2N)$ [99, p 158].)

Another version transforms the matrix $\tilde{H}_l$ into an upper triangular matrix using plane rotations/Givens rotation matrices, to find the residual $\|\mathbf{b} - A\mathbf{x}_l\|$ for each iterate $\mathbf{x}_l$ and hence give a stopping criteria [99, §6.5.3], with *"virtually no additional arithmetic operations"* [99, p 170]. The version of GMRES that we use (the MATLAB implementation) in later numerical experiments (see §5) involves the key elements of both the Householder orthogonalization and rotation matrices versions of GMRES described in this and the previous paragraph.

Another variant is 'restarted GMRES', where information on previous iterations is discarded after l iterations and the algorithm is restarted using $\boldsymbol{x}_0 := \boldsymbol{x}_l$ [99, §6.5.5]. However this method can stagnate if the matrix is not positive definite [99, §6.5.5]. We do not consider the restarted version elsewhere in this thesis.

An upper bound on the cost in terms of floating point operations of the GMRES algorithm for all the variants that we have mentioned, with the exception of restarted GMRES, is as follows. Let $N_Z(A)$ be the number of non-zero entries in A, then l steps of the Arnoldi method requires l matrix-vector products (see Algorithm 1.5.2 Step 3), the cost of which is $\mathcal{O}(lN_Z(A))$ [99, p 160]. For each iteration, the steps in the Gram-Schmidt process (see Algorithm 1.5.2 Steps 5-7) costs $\mathcal{O}(jN)$ operations, so that the total after l iterations is $\mathcal{O}(l^2N)$ [99, p 160].

These costs dominate the costs of the algorithm (Algorithm 1.5.2 Steps 8 and 9 are $\mathcal{O}(lN)$ and Algorithm 1.5.2 Steps 11 and 12 are $\mathcal{O}(l^3)$ by the cost of least squares). Therefore the overall cost of Algorithm 1.5.2 is $\mathcal{O}(lN_Z(A)+l^2N)$ [99, p 160].

An upper bound on the storage cost of the GMRES algorithm for all the variants that we have mentioned, with the exception of restarted GMRES, is as follows. There are l basis vectors $\boldsymbol{v}_i$ of length N and also the vectors $\boldsymbol{b}, \boldsymbol{x}_l$, and $\boldsymbol{w}_j$ of length N to be stored, also the Hessenberg matrix $\tilde{H}_l$ of size $\frac{l^2}{2}$. Therefore the overall storage cost is

$$\mathcal{O}(lN + l^2). \tag{1.29}$$

As long as l, the number of iterations is small (ie. $l \ll n$), these costs are $\mathcal{O}(N)$ (depending upon the value of $N_Z(A)$). However for larger values of l the cost increases considerably. This increasing cost is an important consideration for use of GMRES: the Conjugate Gradient method for Hermitian matrices has cost which grows much slower in l.

1.5.2.2 Choice of Iterative Method

The simplest Krylov subspace algorithm is arguably the Conjugate Gradient method. However this method only works well in the case of Hermitian, positive-definite matrices. We recall from Remarks 1.1.4 and 1.4.6 that matrices arising from discretisations of scattering boundary value problems with corresponding approximations to the Sommerfeld radiation condition are non-self-adjoint, i.e., they are not Hermitian (in fact they are symmetric but complex-valued). Therefore a different Krylov subspace method needs to be used. Here we consider the choices of GMRES and Bi-CGSTAB, that work for non-Hermitian and non-symmetric problems.

The theory of GMRES is well established, for example GMRES does not break down except when the residual is zero [99, §6.5.4], it minimises the residual norm over the space $\mathbf{x}_0 + K_l(A, r_0)$ [99, p 164] and must converge within N steps [99, p 172] (assuming exact arithmetic). (Preconditioners are used to minimise the number of iterations l, as outlined in the next section.) In contrast, there is no convergence theory of comparable rigour for Bi-CGSTAB [74, p50].

A comparison of the costs of the two methods can also factor into a decision

between the methods. A fundamental difference between the two methods is that GMRES works out the next iteration based on vectors from all the previous iterates in the matrix $\tilde{H}_l$ (see Algorithm 1.5.2) of size of $(l+1) \times l$, whereas Bi-CGSTAB is a recurrence method and has fixed storage costs per iteration (in fact the memory cost is only that of a fixed number of vectors at each iteration so that the cost is $\mathcal{O}(N)$ and does not therefore change with l [74, p50]). Therefore, when iteration numbers l are higher, the memory costs of GMRES can become prohibitive, whilst those for Bi-CGSTAB remain manageable [74, p50].

With regards the computational costs, GMRES requires only one matrix vector product involving A per iteration (see Algorithm 1.5.2), whereas Bi-CGSTAB requires two [74, p50], so that the computational cost per iteration may be lower for GMRES if it is expensive to multiply by A and the number of iterations is small and in these circumstances GMRES is the method of choice to minimise cost [58, p92]. However, when iteration numbers l are large, the computational costs of GMRES increase dramatically; recall that the computational cost is $\mathcal{O}(lN_Z(A)+l^2N)$. Therefore, if the number of GMRES iterations is high, another recurrence method like Bi-CGSTAB is more efficient in terms of computational cost, see [74, p50] [58, p92].

Whilst we can easily estimate the cost per iteration for any a particular matrix, in practice, the overall cost involves the number of iterations and very little can be said about the convergence rate of even GMRES (indeed for a general matrix, any non-increasing convergence curve is possible, including no convergence until the last iteration, when the residual drops to 0 [59]). As Bi-CGSTAB uses an iteration of restarted GMRES within its algorithm, in a rather ad-hoc way, even less can be said about the convergence theory for Bi-CGSTAB. (For example, note the absence of convergence theory in the book by van der Vorst, the creator of Bi-CGSTAB, in [111, §7]). For matrices with particular properties, a limited amount of convergence theory exists for GMRES, we look at the Elman estimate in §1.6 and §1.9.

Our small demonstrative experiments using GMRES do not encounter significant limiting memory problems (presumably partly because the preconditioners (see §1.6) that we use in most cases reduce the number of iteration counts to a relatively small number of iterations, see §5, though also because our system size N is limited). Bi-CGSTAB may be more cost-efficient in certain circumstances,

such as when iteration counts are high, but as its convergence isn't guaranteed we choose to use GMRES in our experiments in the first instance, due to its more well established convergence theory and better guaranteed convergence properties.

1.6 The Need for Preconditioners for Helmholtz Problems

Simply applying an iterative method to a linear system usually does not result in a cost-effective solution of the problem as, depending upon the conditioning of the matrix in the linear system A, it may take a large number of iterates $\mathbf{x}_i$ to converge to the true solution. To solve this problem, a preconditioner is usually applied, seeking to reduce the number of iterations needed before the method converges.

At the simplest level, a left-preconditioning matrix B, should have the following properties:

i) solving $B^{-1}A\mathbf{x} = B^{-1}\mathbf{b}$ using an iterative method should require fewer iterations than simply applying the iterative method to the original system $A\boldsymbol{x} = \boldsymbol{b}$,

and

ii) it should be cheap to find and apply the action of multiplying by B^{-1} (the matrix B or its inverse B^{-1} need not be found explicitly).

Specialised preconditioners are required for solving Helmholtz problems with large wavenumber k with Krylov subspace methods, as without them the number of iterations, and hence the cost, increases dramatically, so as to be prohibitive [38].

The reasons that GMRES performs poorly for Helmholtz problems are subtle and there is a comprehensive review of problems using various iterative methods to solve Helmholtz problems in [38]. For discretisations of Laplace-type problems, linear systems can be solved efficiently using multi-grid methods, see for example [62]. However, due to the highly oscillatory nature of Helmholtz problems, general multi-grid techniques are not effective for solving Helmholtz problems [38].

Lack of coercivity of standard variational formulations of Helmholtz problems is commonly cited as the reason iterative solvers encounter problems in the solution of Helmholtz problems, see for example [19, p195]. The standard variational formulation of the Helmholtz IIP is not coercive [28]. For the sesquilinear form (1.22) arising in the interior impedance problem's variational formulation, [105, Lemma 6.5] says that for k sufficiently large there exists a $v \in H^1(\Omega)$ with $a(v, v) = 0$. If the relevant finite element formulation for this sesquilinear form is used, with a sufficiently fine grid discretisation, the matrix would therefore have the origin in its field of values.

When the problem is coercive, 0 is not contained in the field of values. In this case, the Elman estimate can be applied, and provides an explicit relationship between the field of values of a matrix and convergence of GMRES.

Theorem 1.6.1. ***(Elman estimate)*** *Let A be a matrix with $0 \notin W(A)$, where $W(A) := \{(A\boldsymbol{u}, \boldsymbol{u}) : \boldsymbol{u} \in \mathbb{C}^N,\ \|\boldsymbol{u}\|_2 = 1\}$ is the numerical range or field of values. Let β be defined such that*

$$\cos\beta = \frac{\operatorname{dist}(0, W(A))}{\|A\|_2}.$$

Then, if the matrix equation $A\boldsymbol{u} = \boldsymbol{f}$ is solved using GMRES, for $m \in \mathbb{N}$, the GMRES residual $r_m := A\boldsymbol{u}_m - \boldsymbol{f}$ satisfies

$$\frac{\|r_m\|_2}{\|r_0\|_2} \leq \sin^m\beta. \tag{1.30}$$

The bound (1.30) *was originally proved in [31] (see also [30, Theorem 3.3]) and appears in the form above in [7, Equation 1.2].*

This estimate implies that the larger the distance between the field of values and the origin, the better the rate of convergence of the residual to 0. In particular, if the field of values contains zero, i.e. the matrix is indefinite, the Elman estimate cannot be applied. Note that zero being in the field of values does not necessarily mean that GMRES will perform poorly (since the conditions in the Elman estimate are sufficient but not necessary) but given the lack of any other rigorous theory, the above gives us some insight into the reason that preconditioners are needed in the case of large k.

The above discussion has only been for the FEM discretisation of the interior impedance problem and not for other variations of Helmholtz problems. However, given that the impedance boundary condition can been seen as just another way of approximating the SRC (see §1.3), one expects the same large iteration counts to be encountered in discretisations of Helmholtz problems involving PML and other artificial boundary conditions used to approximate the SRC (see §1.3). Indeed the same large iteration counts are seen for other ways of discretising the problem [34, 35, 42].

However, the lack of coercivity is not the sole reason behind the difficulties iterative methods encounter when solving Helmholtz problems. In [86], Moiola and Spence create a formulation with a coercive bilinear form. Since the formulation is coercive, there exists a minimum distance between the field of values and the origin [28, p5]. However, the iteration counts for this formulation still grow with k [28, Figure 7], showing that there are more subtle problems as well as the lack of coercivity in standard variational formulations.

Therefore it is natural to seek a preconditioner for these types of problems with the goal of reducing the number of GMRES iterations. Indeed, an ideal case would be one where the number of iterations and other storage and memory costs of the overall calculation using the iterative method are bounded independent of k.

1.7 Sweeping Preconditioners

1.7.1 Introduction to Sweeping Preconditioners

There are lots of preconditioners for Helmholtz problems. We only focus on sweeping-type preconditioners, first developed by Engquist and Ying [34, 35].

A significant recent development is that Gander and Zhang have described the sweeping-type classes of preconditioners and several other classes of preconditioners using a common framework based on optimised-Schwarz domain-decomposition methods [44]. Classes of preconditioners thus described are sweeping [34,35], source transfer [15], single-layer potentials [106], polarised traces [115] and optimised-Schwarz methods [45] (all references are a sample of the literature for each type).

Sweeping-type preconditioners have very good iteration counts with GMRES, showing only a weak dependence upon the wavenumber k in [34, 35]. We go through key idea of formulating sweeping-type preconditioners in the next section §1.7.2 and later give a detailed formulation of them in §2.

Some examples of the particular formulation of sweeping preconditioner given in §1.7.2 and §2, where the key step is approximating Schur complements, are [34, 35, 46, 114]. More details on how the methods of approximating the Schur complements of interest to this thesis are given at the end of §2.

There are many related formulations to the formulation in §1.7.2 and §2, for example in [80] Liu and Ying solve the various subdomain problems, but send solutions from the subdomain boundaries to up and down between the subdomains, adding up the solutions as go through the subdomains to make the method efficient.

1.7.2 Key idea

A key idea of this thesis is that the action of Schur complement matrices arising in the formulation of sweeping preconditioners is equivalent to solving certain Helmholtz boundary value problems (BVPs), the solution operators for which can be proved to be low-rank.

To demonstrate this key idea we look at an illustrative problem based on only two subdomains. This problem is posed on the half-plane with zero-Dirichlet boundary condition. The problem is truncated to a small domain of interest (Ω) with Perfectly Matched Layers (PMLs) on 3 sides as in Figure 1-6. (For details about PML, recall the discussion in §1.3.) We assume that the truncated problem is discretised on a tensor product grid, the nodes of which are divided between two non-overlapping subdomains Ω_1 and Ω_2 (see Figure 1-6).

Applying a finite difference or finite element method to the BVP creates a linear system

$$A\boldsymbol{u} = \boldsymbol{f}.$$

We block this system according to the decomposition of the domain $\Omega = \Omega_1 \cup \Omega_2$

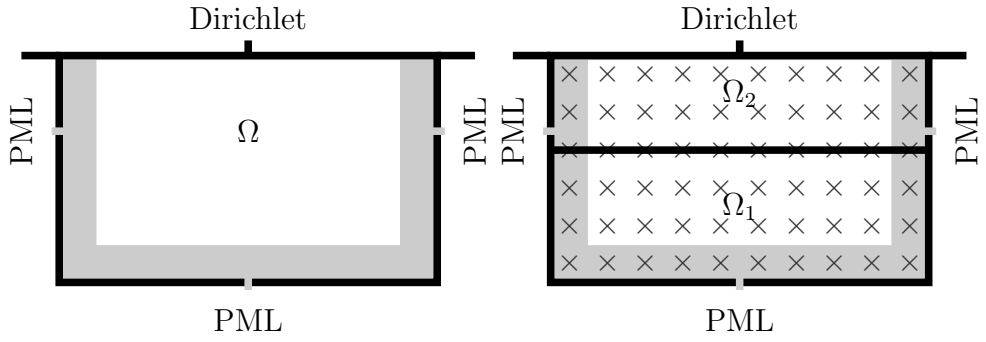


Figure 1-6: Domains of illustrative problem. Left: Ω is the region of interest below the zero-Dirichlet condition on the upper boundary of the half-plane problem with PMLs on the three other sides. Right: Ω is discretised and subdivided into Ω_1 and Ω_2, with the boundary along a line of nodes from the tensor product grid.

in the natural way to obtain

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

(Note that $A_{1,2}$ from the natural blocking of the system represents the interaction with rows corresponding to nodes in Ω_1 and columns corresponding to nodes in Ω_2. In particular, in this decomposition the nodes on boundaries of the domains Ω_m are assumed to be in the blocks corresponding to the lower domain.) We define the Schur complement matrix

$$S := A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}, \tag{1.31}$$

and the modified source

$$\tilde{f}_2 := f_2 - A_{2,1}A_{1,1}^{-1}f_1, \tag{1.32}$$

so that after one step of block elimination the system becomes

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & S \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ \tilde{f}_2 \end{bmatrix}. \tag{1.33}$$

The formulation of a sweeping preconditioner for our simple illustrative problem has 2 basic ideas underlying it.

1st idea Sweeping

The reason for creating subdomains and blocking the linear system is to solve the whole problem by the method of solving a series of smaller, easier subproblems sequentially. This method takes the following form: assuming we can invert $A_{1,1}$, or cheaply approximate the action of multiplying by $A_{1,1}^{-1}$ (without necessarily finding $A_{1,1}$ explicitly), we can solve on Ω_2 via (1.31), (1.32) and (1.33) and finally solve for u_1 on Ω_1. (Multiplying by $A_{1,1}^{-1}$ is in some sense solving the subproblem on Ω_1, as we see in the section 2nd idea detail.) This method forms a sweeping algorithm as follows:

$$\begin{aligned} &\text{Solve on } \Omega_1 \\ &\qquad\qquad \downarrow \text{ upward sweep} \\ &\text{Solve on } \Omega_2 \\ &\qquad\qquad \downarrow \text{ downward sweep} \\ &\text{Solve on } \Omega_1. \end{aligned}$$

2nd idea Approximate multiplying by S^{-1}.

For the solve on Ω_2 we need to invert S (or cheaply approximate the action of multiplying by S^{-1}). We have an indication that finding such a cheap approximation is possible due to the fact that

$$\text{action of } S^{-1} \approx \text{action of a Helmholtz solution operator that can be shown to be low-rank.}$$

This fact is the key theoretical idea behind Engquist and Ying's sweeping preconditioner in [34] and as it is also key for this thesis we discuss it in detail in §2.2.3. This thesis gives new results on the low-rank expansion of the Helmholtz solution operator and works out consequences of these new results for sweeping preconditioners. To get to the context of these results we look in more detail at the 2 ideas.

1st idea detail

We present the full sweeping algorithm for our illustrative systems as follows.

Algorithm 1.7.1 (Illustrative Sweeping Algorithm)**.**
Upward Sweep

1. Compute $v_1 = A_{1,1}^{-1} f_1$ *Zero-Dirichlet solve on* Ω_1

2. *Compute* $\tilde{f}_2 = f_2 - A_{2,1}v_1$ *Modified source on* Ω_2

Downward Sweep

3. *Compute* $u_2 = S^{-1}\tilde{f}_2$ *Zero-Dirichlet solve on* Ω_2

4. *Compute* $\tilde{f}_1 := f_1 - A_{1,2}u_2$ *Modified source on* Ω_1

5. *Compute* $u_1 = A_{1,1}^{-1}\tilde{f}_1$ *Zero-Dirichlet solve on* Ω_1

To see the first that step 1 is a zero-Dirichlet solve on Ω_1, observe that if we were to chop Ω_2 out of the problem (or equivalently truncate the problem to Ω_1) and discretise, the linear system would consist simply of $A_{1,1}v_1 = f_1$, for some $v_1 \neq u_1$. We note that $v_1 \neq u_1$, because the problem has been altered: by truncating and considering only Ω_1 we have effectively set the value of the solution to zero everywhere in Ω_2, in particular along the row of nodes above the boundary of Ω_1, creating a new zero-Dirichlet boundary above Ω_1. Therefore inverting $A_{1,1}$ is in some sense equivalent to solving the half-plane Helmholtz problem with zero-Dirichlet condition on Ω_1. A more detailed explanation of this, in a more general setting with several subdomains, can be found in §2.2.11.

Next, step 2 is a modified source as the $-A_{2,1}v_1$ term is transferring the effect of the source in Ω_1 (as the wave would) to incorporate it into the source in Ω_2, creating $\tilde{f}_2$.

Due to the zero-Dirichlet condition at the top of the original problem we see step 3 is a zero-Dirichlet solve on Ω_2. Finally the description for steps 4 and 5 follow as in steps 2 and 1 respectively.

In practice to form a preconditioner the steps in Algorithm 1.7.1 are not all completed exactly and S^{-1} is not necessarily found explicitly, rather a cheap way to approximate the action of multiplying by S^{-1} is found. Assuming the same can be done for $A_{1,1}^{-1}$, Algorithm 1.7.1 is then a relatively low cost algorithm. (To see this observe that $A_{1,2}$ and $A_{2,1}$ are diagonal matrices and all the other operations we have not already considered are vector operations.)

The theoretical justification for it to be possible to cheaply approximate the action of S^{-1} is found in the 2nd idea which we look at in more detail now.

2nd idea detail

To understand the action of S^{-1} we do the same block elimination as we used to form (1.33), only this time for an artificial – but related – problem: this

problem has source defined only on Ω_2. So we solve

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ g \end{bmatrix},$$

becomes

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & S \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ g \end{bmatrix},$$

where S is precisely the Schur complement from (1.31) (but as there is no source in Ω_1 there isn't an equivalent modified source (1.32)). We see that $w_2 = S^{-1}g$; in other words the action of S^{-1} is to solve the Helmholtz problem on the whole of Ω but with the source having support only on Ω_2 and observing the solution only on Ω_2.

Consequently, the action of S^{-1} is associated with a Helmholtz solution operator, an operator that crucially is low-rank (we look at low-rank properties in §1.8 and we discuss the action of S^{-1} in §2.2.3). As the associated Helmholtz solution operator is low-rank, the matrix S^{-1} admits "good" low-rank approximations itself. This is what allows the S^{-1} matrices to be cheaply approximated (by, for example, $\mathcal{H}$-matrices, see §1.8.3) and makes the sweeping preconditioner computationally attractive.

In practice, the problem is divided into many more than two subdomains. We give a detailed formulation of a many subdomain sweeping preconditioner in §2.

1.8 Low-Rank Approximations

1.8.1 Low-Rank Approximations of Green's Functions

Definition 1.8.1. ***(Low-rank separable expansion)*** *Let $X, Y \subset \mathbb{R}^d$ be subsets. A function $\varkappa(x, y)$, $x \in X$, $y \in Y$, has a low-rank separable expansion if there exists p and functions $\{\phi_j, \psi_j\}_{j=1}^p$ such that*

$$\left| \varkappa(x, y) - \sum_{j=1}^{p} \phi_j(x)\psi_j(y) \right| < \varepsilon, \tag{1.34}$$

for all $x \in X$, $y \in Y$, where the rank p is small.

(For variations on this definition of the low-rank separable expansion, see, for example [33, Definition 2.1] and [6, Definition 3.8].)

The expansion is separable because the ϕ and ψ functions in the expansion are functions of only x or y, separating the dependence on the variables. A function with a low-rank separable expansion is sometimes called 'degenerate' [6, p117].

There are two main contexts for considering low-rank separable expansions of fundamental solutions/Green's functions. The first is integral methods (as mentioned in §1.2) where the kernel is related to the fundamental solution (i.e., it is either the fundamental solution itself or a derivative of it). The second is domain based methods, either sweeping methods (see the 2nd key idea in §1.7.2) or direct solvers (as mentioned in §1.5.1). Where low-rank separable expansions for functions exist, they can often be exploited by numerical methods to efficiently solve the associated problems (see §1.8.2).

Definition 1.8.2. ***(Asymptotically Smooth)*** *[63, Definition 4.14] Let $X, Y \subset \mathbb{R}^d$ be subsets. Let the function $\varkappa(x,y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ be arbitrarily often differentiable for all $x \in X$ and $y \in Y$ with $x \neq y$. Then $\varkappa(x,y)$ is asymptotically smooth in $X \times Y$ if*

$$\left|\partial_x^\alpha \partial_y^\beta \ \varkappa(x,y)\right| \leq c_{\text{as}}(\alpha+\beta)|x-y|^{-|\alpha|-|\beta|-\sigma}, \tag{1.35}$$

for $x \in X$, $y \in Y$, $x \neq y$, $\alpha, \beta \in \mathbb{N}_0^d$, $\alpha + \beta \neq 0$, holds for some $\sigma \in \mathbb{R}$ and

$$c_{\text{as}}(\nu) = C\nu!|\nu|^\rho \gamma^{|\nu|},$$

where $\nu \in \mathbb{N}_0^d$ and C, γ and ρ are constants.

(There are many slight variants on the definition of asymptotically smooth in the literature, for example an alternative definition allowing for $|\varkappa|$ on the right hand side of (1.35) as well is found in [6, Definition 3.2].)

If a function $\varkappa(x,y)$'s derivatives are controlled, in the sense that $\varkappa$ is asymptotically smooth, there are theorems that show the interpolant of $\varkappa(x,y)$ in either x or y approximates $\varkappa$ with the quality of the approximation improving algebraically with the order of interpolation, see [63, Theorem 4.22] or [6, Lemmas 3.16-7 and Theorem 3.18]. These theorems form the basis of finding low-rank separable expansions to such asymptotically smooth functions $\varkappa$. (We use this

theory as part of finding a separable expansion for the Hankel function (1.3) in §3.4.4.)

The fundamental solution of the Laplace equation is asymptotically smooth and therefore admits a low-rank separable expansion [63, Appendix E.1], an example of "good" low-rank properties.

However, the derivatives of the fundamental solutions of the Helmholtz equation grow with k. Consequently, on general domains, Helmholtz fundamental solutions have "poor" low-rank properties in the large-k limit. Since the size of their derivatives increase with k, although the fundamental solutions are still asymptotically smooth, the parameters of asymptotic smoothness depend on k and the ranks p of their "low"-rank separable expansions therefore also increase with k. Indeed, Engquist and Zhao show that for a general pair of disjoint compact domains in 3D, the lower bound on the rank p is k^2 [36, Example 4.1] and for a general pair disjoint compact surfaces (a common occurrence in boundary integral methods), in 3D, the lower bound on the rank p is k^2 [36, Example 4.1].

However, if some directionality and distance is imposed on the domains X and Y, it is possible to get "good" low-rank separable expansions of Helmholtz fundamental solutions in the large-k limit. In 3D, Engquist and Zhao show that for two collinear line segments [36, §4.2 Example 1] and two 'collinear' cylinders [36, §4.2 Example 2] the upper bound on the rank p does not depend on k, but only on $\log(\varepsilon)$. (Previously, Engquist and Ying [33], as part of the development of their directional multilevel algorithm for solving N-point problems with highly oscillatory kernels, looked at low-rank separable expansions in distant domains within certain cone angles.) In 2D Rokhlin and Martinsson [82] have a low-rank result on long, thin 'collinear' domains of particular interest in this thesis, see the statement in Theorem 2.2.22.

Sometimes, in low-rank results for Helmholtz fundamental solutions, the diameter and distance of the domains X and Y are dependent upon k. In 3D, when X and Y are spheres, Delamotte et. al. showed that the rank of the Helmholtz fundamental solution eventually depends at most linearly on k, provided the spheres satisfy the Fresnel condition, which means that the ratio of the diameters and distance of the spheres must be k dependent [25, 26]. Börm and Melenk similarly obtain low-rank approximations to the Helmholtz fundamental solutions in 2D and 3D, that depend on the ratio of the diameters and distances of

axis-parallel boxes X and Y being dependent upon k [11, Theorem 3.13, Lemma 4.2 with conditions (3.20)].

Conditions that impose the required relationship between the diameter and distance or directionality and distance of the domains for low-rank results to be obtained are generally called admissibility conditions. Our low-rank results contain an admissibility condition ($\eta \text{dist}(D_1, D_2) > \text{diam}(D_1, D_2)$ for some domains D_1, D_2 and some constant $\eta > 0$, see for example Theorem 3.2.3) and we look at particular cases of weak and strong admissibility in §4.2.2 (see in particular Definitions 4.2.6 and 4.2.6).

1.8.2 Low-Rank Approximation Methods

Due to the low-rank properties of various functions discussed in §1.8.1, discretisations of these functions inherit the property that they readily admit good-quality low-rank approximations.

Perhaps the most famous method that exploits low-rank properties of kernel functions is the Fast Multipole Method. First introduced in [60], it is now used to perform fast matrix-vector multiplication of discrete integral operators (for example those arising in boundary integral methods mentioned in §1.2), see for example [73]. Indeed, most of the low-rank approximation methods are in the context of matrix-vector multiplication of discrete integral operators, as this is the situation that most naturally lends itself to low-rank approximation (the discrete integral operators consist of dense matrices that, depending on the operator, inherit the property that they readily admit good-quality low-rank approximation). Other examples of such methods include panel-clustering [65] and matrix compression using wavelets [24].

Another method that exploits these low rank properties is the method for the direct inversion of integral operators [81, 82] developed by Rokhlin and Martinsson. There are many others.

In this thesis, we are most concerned with $\mathcal{H}$-matrices [56, 57, 61], which allow for not only efficient matrix-vector multiplication of matrices that have low-rank properties, but efficient versions of many other matrix operations as well.

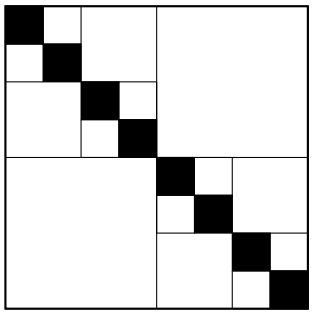


Figure 1-7: The admissible block structure of a $\mathcal{H}$-matrix, white off-diagonal blocks are stored in low-rank factorised form and only black diagonal blocks are stored densely or in full.

1.8.3 Overview of $\mathcal{H}$-Matrices

In this thesis we are especially concerned with $\mathcal{H}$-matrix approximations of Schur complement matrices arising in sweeping preconditioners (see §1.7). $\mathcal{H}$-Matrices (also called Hierarchical Matrices) were proposed by Hackbusch et al. [56, 57, 61]. A $\mathcal{H}$-Matrix is an approximation to a matrix that can be stored and manipulated at a lower cost than the matrix it approximates. A $\mathcal{H}$-Matrix has various off-diagonal blocks, called admissible blocks, stored in a low-rank factorised form, see Figure 1-7. For example an off-diagonal block of a $\mathcal{H}$-Matrix of size $n' \times m'$, is approximated by the factorisation UV^T with matrices $U \in \mathbb{C}^{n' \times R}$ and $V \in \mathbb{C}^{m' \times R}$, for some $R \ll \min\{n', m'\}$, see Figure 1-8.

Creating the decomposition of the matrix into admissible blocks to be stored in low-rank factorised form and inadmissible blocks to be stored densely is a non-trivial task. The partition of the matrix is characterised by cluster trees. There are many variants of $\mathcal{H}$-matrices, we mention 'standard' $\mathcal{H}$-matrices which we consider in this thesis [63] and $\mathcal{H}^2$ matrices or Hierachically Semi-Separable matrices (see for example [14, 64]), of incidental interest to us. In §4.2.2 we recall a derivation of strongly and weakly admissible 'standard' $\mathcal{H}$-matrices.

$\mathcal{H}$-matrices are most commonly used in the context of boundary integral methods, rather than volume or domain-based discretisation methods. However, the geometric regions associated with the Schur complement matrices are subdomains, not boundaries, of the problem, so we use $\mathcal{H}$-matrices in the less common context of volume or domain-based discretisation.

The Hierarchical Matrix Framework (HMF) is the framework for efficiently

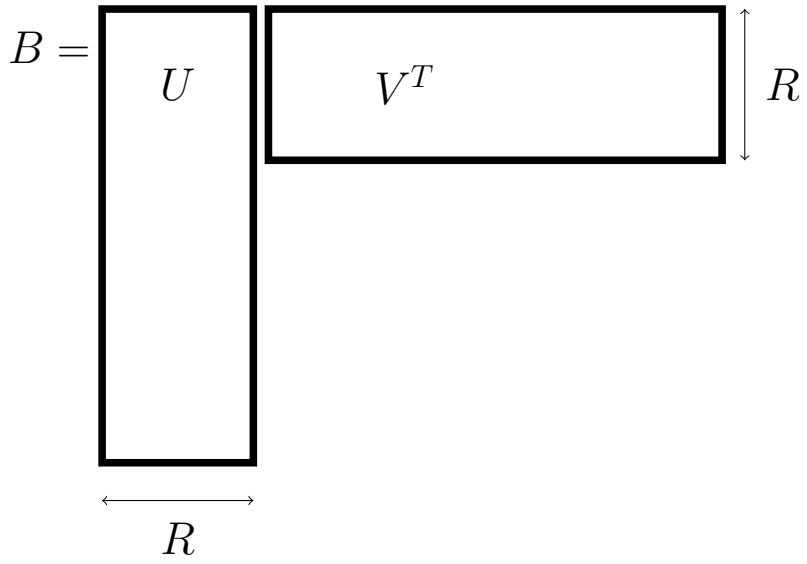


Figure 1-8: How an off-diagonal block B is approximated in low-rank form in a $\mathcal{H}$-Matrix.

creating and manipulating $\mathcal{H}$-Matrices. We note that creating $\mathcal{H}$-matrix approximations efficiently is only part of what is needed to use $\mathcal{H}$-matrices in practice. Algorithms are needed to perform matrix operations (for example matrix-vector multiplication, matrix-matrix addition and multiplication and matrix inverse) using $\mathcal{H}$-matrices cost efficiently, by taking advantage of the low-rank structure. For complete details about how these calculations are performed in practice, we refer the reader to [56, 57, 61, 63]. In this thesis we describe how a limited number of matrix operations are performed, for one set of our numerical experiments conducted with $\mathcal{H}$-matrices, in Appendix A.

1.9 Preconditioners for Helmholtz Problems with Absorption

1.9.1 Different Conventions for Adding Absorption

Before discussing preconditioners with absorption, we first fix some notation.

We can add absorption to the Helmholtz operator in two ways

1) $k^2 \mapsto k^2 + \mathrm{i}\alpha$,

2) $k \mapsto k_R + \mathrm{i}k_I$.

The literature on the preconditioner with absorption usually follows the first convention and we use the first convention in §1.9.2.3. However we wish to highlight that in later chapters we follow the second convention, writing k as $k_R + \mathrm{i}k_I$. (We use this convention for convenience as we work mainly with k rather than k^2.)

We give some illustrative examples to make relationship between the two conventions clear.

Lemma 1.9.1. *i) If $\alpha = k^2$ (i.e. $k^2 + \mathrm{i}k^2 = (k_R + \mathrm{i}k_I)^2$), then*

$$k_R = \sqrt{\frac{1+\sqrt{2}}{2}}\, k \qquad \textit{and} \qquad k_I = \frac{k}{2\sqrt{\frac{1+\sqrt{2}}{2}}}.$$

ii)If $\alpha = k$ (i.e. $k^2 + \mathrm{i}k = (k_R + \mathrm{i}k_I)^2$), then

$$k_R = \frac{1}{\sqrt{2}}\sqrt{1+\sqrt{1+\frac{1}{k^2}}}\, k \qquad \textit{and} \qquad k_I = \frac{1}{\sqrt{2}}\frac{1}{\sqrt{1+\sqrt{1+\frac{1}{k^2}}}}.$$

Proof. Substitute the values of k_R and k_I into $(k_R + \mathrm{i}k_I)^2 = k_R^2 - k_I^2 + 2\mathrm{i}k_R k_I$. □

Therefore, despite being written in the different conventions, the condition $\alpha \sim k$ is equivalent to the condition $k_I \sim 1$ and the condition $\alpha \sim k^2$ is equivalent to the condition $k_I \sim k$.

1.9.2 Preconditioners with Absorption

1.9.2.1 Basic Principle of Adding Absorption

An example of a Helmholtz problem with absorption is

$$\Delta_x u(x) + (k^2 + \mathrm{i}\alpha)u(x) = -f(x), \quad \text{in } \mathbb{R}^2,$$

where f has compact support, with the Sommerfeld radiation condition

$$\frac{x}{\|x\|}.\nabla u(x) - \mathrm{i}\sqrt{k^2 + \mathrm{i}\alpha}\, u(x) = o\left(\frac{1}{\|x\|}\right), \quad \text{as } \|x\| \to \infty.$$

The idea of adding absorption is that the oscillations of the solution are damped. We can see that the solutions to the problems are less oscillatory by looking at the fundamental solutions: in both 2D and 3D the fundamental solution of the problem with absorption includes the oscillatory factor

$$\exp(\mathrm{i}k\|x-y\|) = \exp(\mathrm{i}\sqrt{k^2+\mathrm{i}\alpha}\|x-y\|)$$

$$= (\exp(\mathrm{i}(k_R + \mathrm{i}k_I)\|x-y\|) = \exp((k_R\mathrm{i} - k_I)\|x-y\|)\,. \tag{1.36}$$

To see the oscillatory factor in the fundamental solutions, recall the fundamental solutions are given in §1.1.1.1. The oscillatory factor is explicitly in the expression for the fundamental solution in 3D. In 2D the factor can be seen in this expansion of the Hankel function:

$$H_0^{(1)}(z) = -\frac{2\mathrm{i}}{\pi}\exp(\mathrm{i}z)\int_0^\infty \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}}\,\mathrm{d}t, \quad \text{for } 0 < \mathrm{Re}(z) < \infty; \tag{1.37}$$

this expansion can be found in [75, §4.1.2 (4.19)] due to [90, §7.13.3 (13.07)] for $z \in \mathbb{R}^+$ (the positive real line, not including zero). We look at the expansion (1.37) in §3.4.2 and find that it is also valid for complex argument z, hence we state the range in (1.37) as $0 < \mathrm{Re}(z) < \infty$. Therefore as the imaginary part of k increases it damps the oscillations of $H_0^{(1)}$ that increase with the real part of k, as in (1.36).

In Chapter 3 we investigate the damping effect on the low-rank properties of the 2D fundamental solution.

Definition 1.9.2. ***(Matrix** A**)*** *We discretise a Helmholtz problem with a finite element method (in Chapter 4) or finite difference method (in Chapter 5) creating the linear system*

$$A\boldsymbol{u} = \boldsymbol{f}.$$

Definition 1.9.3. ***(Matrix** A_{abs}**)*** *We add absorption to the wavenumber of a Helmholtz problem so that $k(x) \in \mathbb{C}$. We discretise the new model problem with*

a finite element or finite difference method creating the linear system

$$A_{\text{abs}}\boldsymbol{u}_{\text{abs}} = \boldsymbol{f}.$$

Definition 1.9.4. ***(Matrix $\tilde{A}_{\text{abs}}$ and Preconditioned Systems)*** *An approximation of A_{abs}^{-1}, denoted by $\tilde{A}_{\text{abs}}^{-1}$, is the preconditioner with absorption. The preconditioned system is then*

$$\tilde{A}_{\text{abs}}^{-1}A\boldsymbol{u} = \tilde{A}_{\text{abs}}^{-1}\boldsymbol{f}. \tag{1.38}$$

1.9.2.2 Theory of Adding Absorption

To understand why constructing the preconditioner from a different problem is more effective, we follow parts of the heuristic argument in [42].

We begin by looking at Corollary 1.9.5 of the Elman Estimate for our preconditioned matrix A_{abs}^{-1}. Corollary 1.9.5 gives an estimate for β, which we recall from (1.30) controls the rate of convergence to 0 of the GMRES residual, in terms of $\|I - \tilde{A}_{\text{abs}}^{-1}A\|_2$ (where I is the identity matrix).

Corollary 1.9.5. *[42, Corollary 1.9] If $\|I - \tilde{A}_{\text{abs}}^{-1}A\|_2 \leq \sigma < 1$ and if β is as in (1.30) for $A := \tilde{A}_{\text{abs}}^{-1}A$ and $\boldsymbol{f} := \tilde{A}_{\text{abs}}^{-1}\boldsymbol{f}$, then*

$$\cos\beta \geq \frac{1-\sigma}{1+\sigma} \quad \textit{and} \quad \sin\beta \leq \frac{2\sqrt{\sigma}}{(1+\sigma)^2}.$$

We recall from (1.30) that if $\sin\beta$ is small the GMRES residual converges to 0 quickly, so this corollary tells us that GMRES works well if $\|I - \tilde{A}_{\text{abs}}^{-1}A\|_2$ is sufficiently small. We then observe that

$$I - \tilde{A}_{\text{abs}}^{-1}A = I - \tilde{A}_{\text{abs}}^{-1}A_{\text{abs}} + \tilde{A}_{\text{abs}}^{-1}A_{\text{abs}}(I - A_{\text{abs}}^{-1}A),$$

so $\|I - \tilde{A}_{\text{abs}}^{-1}A\|_2$ is sufficiently small if $\|I - \tilde{A}_{\text{abs}}^{-1}A_{\text{abs}}\|_2$ and $\|I - A_{\text{abs}}^{-1}A\|_2$ are sufficiently small. We express these conditions as

P1) $\tilde{A}_{\text{abs}}^{-1}$ is a good preconditioner for A_{abs},

P2) A_{abs}^{-1} is a good preconditioner for A.

For **P1)** we need α to be large. For **P2)** we need α to be small.

Regarding **P1)**, for a general preconditioner $\tilde{A}^{-1}_{\text{abs}}$, to achieve **P1)** heuristic arguments imply that α needs to be large. This is due to the facts that
1) we recall from §1.6, that the desirable properties for Helmholtz preconditioners are to have k-independent iteration counts and costs, so the difficulties arise when k is large, i.e. for more oscillatory problems.
2) the oscillations of the solution to the problem are increasingly damped as α increases (recall the discussion about the damping factor in equation (1.36) in §1.9.2.1).

For specific preconditioners $\tilde{A}^{-1}_{\text{abs}}$, analyses and numerical experiments have been performed that investigate preconditioning A_{abs} with $\tilde{A}^{-1}_{\text{abs}}$. A recent summary is given in [54, p5]. For multi-grid to converge in a k-independent number of steps $\alpha \sim k^2$ is required [18, 23, 53]. For classical Additive Schwarz domain decomposition preconditioners, in [54] it is shown using the Elman estimate (recall Theorem 1.6.1 in §1.6), that under appropriate conditioning on the domain decomposition, $\tilde{A}^{-1}_{\text{abs}}$ is a good preconditioner for A_{abs} (in that GMRES converges in a k-independent number of iterations) if $\alpha \sim k^2$.

Regarding **P2)**, obviously when $\alpha = 0$, $A^{-1}_{\text{abs}} = A^{-1}$, so A^{-1}_{abs} approximates A^{-1} well; but as α gets larger and the Helmholtz problems the A^{-1}_{abs} matrices originate from become increasingly different to the problems the A^{-1} matrices originate from, we expect the approximation of A^{-1}_{abs} by A^{-1} to get worse.

So overall, to balance the conflicting needs of **P1)** and **P2)** on the value of α, we expect that some, but not too much absorption α, may be of benefit in reducing the number of GMRES iterations.

This argument applies equally well for the IIP or other linear systems, for example those approximating the SRC with PMLs. For the IIP discretised with the FEM, [42, Theorems 1.4 and 1.5] prove the bound $\|I - A^{-1}_{\text{abs}}A\|_2 \lesssim \frac{\alpha}{k}$ (under certain conditions on the problem geometry and discretisation method) i.e. **P2)** is satisfied if α/k is sufficiently small.

We do not present theory for other linear systems, but we expect that any linear systems that are approximations to the same model problem benefit from including absorption in a similar way.

1.9.2.3 Examples of Preconditioners with Absorption

The idea of adding absorption can be used in conjunction with any strategy for preconditioning Helmholtz problems. We recall previous examples of preconditioners with absorption that fall into four main classes, as follows.

The multi-grid method with absorption is known as the Shifted-Laplacian preconditioner [37]. Standard multi-grid methods perform poorly with Helmholtz problems, but when a sufficiently large value of α is included they perform well again [38]. However, for values of α that are sufficiently large for multi-grid to work, the preconditioner is too far away from the problem without absorption to be effective, so that for this preconditioner the requirements of **P1)** and **P2)** are mutually exclusive [17, 38]. Therefore, further adaptations are needed to use this preconditioner for Helmholtz problems without absorption.

Additive Schwarz domain decomposition preconditioners with absorption were constructed and analysed in [53,55]. Analysis of projection operators (rather than fundamental solutions) and numerical experiments show the effectiveness of these preconditioners under certain conditions.

A variant of the Fast Multipole Method and $\mathcal{H}$-matrix approximations of boundary integral operators have been constructed with absorption [40, 73].

Engquist and Ying add absorption to their moving PML sweeping preconditioner [35], this case is of most interest to this thesis and further discussion is contained in §1.9.3.1.

1.9.3 Motivation for Thesis

1.9.3.1 Benefits of Sweeping Preconditioner with Absorption

The experiments of particular interest to this thesis are low-rank experiments with sweeping preconditioners with absorption. As mentioned in §1.9.2.3, Engquist and Ying included a small amount of absorption $k_I = \mathcal{O}(1)$ in their moving PML preconditioner [35]. Shanks [101] conducted experiments with the moving PML preconditioner to determine the optimal amount of absorption for reducing the iteration counts. Of the options $\alpha \in \{0, k, k^2\}$ and $k_I = 1$, $k_I = 1$ was found to be the optimal value, see [101, Tables 5.7-10].

Despite the use of absorption in one variant of the preconditioner, to the best

of the author's knowledge, no work has been done on examining the effect of absorption in the low-rank results that provide underlying motivation for Engquist and Ying's sweeping preconditioner.

1.9.3.2 How Absorption Affects Low-Rank Properties of Helmholtz Fundamental Solution

We saw in §1.7.2 that the sweeping preconditioner depends on low-rank properties of a certain Helmholtz solution operator. We see in §2.2.3 how this solution operator is closely related to the fundamental solution. To understand the benefits of adding absorption to the sweeping preconditioner, it is therefore natural to look at how absorption affects low-rank properties of the fundamental solution and then the related Helmholtz solution operator.

This question about how absorption affects the low-rank properties of the fundamental solution has been partly addressed for large k_I by Banjai and Kachanovska. Indeed, Banjai's result [3, Lemma 5.6] says that for $k \in \mathbb{C}$ with $k_R/k_I \leq C_0$, for some constant $C_0 > 0$, the 3D Helmholtz fundamental solution (1.4) is asymptotically smooth with coefficients depending on C_0 but not depending on k. In the context of the FMM and boundary integral equations, Kachanovska has low-rank theory for Helmholtz fundamental solutions, though this is similarly focussed on large k_I [73, p25]).

We look at how absorption affects low-rank properties of the fundamental solution for smaller values of k_I. These smaller values of k_I are found to be better in practice, recall that we saw that Engquist, Ying and Shanks used $k_I = 1$ in the sweeping preconditioner [35, 101] (see §1.9.3.1). Also that due to the conflicting requirements of **P1)** and **P2)** in §1.9.2.2, we recall that we expect that some, but not too much absorption α, may be of benefit in reducing the number of GMRES iterations.

1.10 Achievements of this Thesis

In Chapter 2 we take an in-depth look at the formulation of a particular type of sweeping preconditioner. The formulation is based on Engquist and Ying's sweeping preconditioner from [34, 35] and is an extension from two subdomains to many subdomains of the key ideas and Algorithm 1.7.1 we saw in §1.7.2. We

particularly focus on explaining the idea that "the action of a Schur complement matrix $S^{-1} \approx$ action of a Helmholtz operator that can be shown to be low-rank", from §1.7.2. We show that the Schur complement matrices arising in the sweeping-preconditioner formulation are approximately equal to matrices that are discretisations of Green's functions operators for a sequence of half-plane problems, like those Green's functions and problems in Definitions 1.1.5, 1.1.6 and 1.1.7.

As mentioned above, the Helmholtz operator associated with the Schur complement matrices can be shown to be low-rank; in §2.2.3.2 we discuss existing results on this. Then in Chapter 3 we prove new low-rank results that cover the case when the wavenumber in the operator is complex, i.e. the problem has some absorption added. Both the existing and new results about the half-plane Green's functions come from the expression (1.10) of the Green's functions in terms of Hankel functions, and from proving appropriate results about the Hankel functions. The motivation for proving these new results about the Hankel and Green's functions is that they can be used to provide underlying theory for the sweeping preconditioner with absorption, and in particular, understand what advantages it has over the sweeping preconditioner without absorption.

A well-known technique for creating low-rank approximations of matrices is the $\mathcal{H}$-matrix framework. In Chapter 4, we use our low-rank results from Chapter 3 to prove results about how well $\mathcal{H}$-matrices can approximate discretisations of the half-plane Green's function, and hence the Schur complement matrices. We then give numerical results showing that the benefits of absorption established in theory are visible in actual $\mathcal{H}$-matrix approximations.

In Chapter 5 we perform experiments on sweeping preconditioner with absorption – both when the problem to be solved contains absorption (in the sense of Definition 1.9.4) and when it does not. Our goal is to investigate whether the benefits due to absorption seen in the $\mathcal{H}$-matrix approximation translate to benefits in the performance of the preconditioner.

Chapter 2

Description of the Sweeping Preconditioner

In this chapter we describe a particular formulation of the sweeping preconditioner (see §1.7), providing the context to expand on the second key idea we saw in §1.7.2:

> "the action of a Schur complement matrix $S^{-1} \approx$
> action of a Helmholtz solution operator that can be shown to be low-rank".

This second key idea is crucial in understanding and motivating the work of this thesis. This chapter also introduces key concepts and notation that we use in the rest of this thesis.

We go through the following stages.

- In §2.1 we describe methods of discretising the model problem, since these methods affect the structure of the matrix of the discretised problem. The structure is of theoretical importance and we use specific discretisations to ensure a certain structure in later theory and numerical experiments. In §2.1.1 we look in detail at the PMLs which approximate the Sommerfeld radiation condition in the model problem. Then we look at the finite difference discretisation and the finite element discretisation in §2.1.2 and §2.1.3 respectively.
- In §2.2 we outline a multi-line sweeping preconditioner (recall sweeping, the

first key idea in §1.7.2). We discussed sweeping preconditioners generally in §1.7, but now we must focus on a particular example of the many different formulations. The first versions of the particular type of sweeping preconditioners we consider are those introduced by Engquist and Ying in [34,35]. Therefore, the sweeping preconditioner formulation we give in this chapter is based upon the preconditioners in these papers. An important difference is that the formulation of the preconditioner we give includes the case when multiple lines of the grid are swept at once, something discussed only briefly in [35]. (We perform numerical experiments with the sweeping preconditioners [34, 35] in §5 and so we give a full description of them there.) We are particularly concerned with two types of matrices arising in the study of the sweeping preconditioner formulation: Schur complement matrices $\mathbb{S}_m^{-1}$ and related matrices $\mathbb{G}^m$ (formed by pointwise evaluation of a certain Green's function, both defined in §2.2.1). In §2.2.2 we see the importance of cheaply approximating the $\mathbb{S}_m^{-1}$ matrices. In §2.2.3 we establish the connection between the $\mathbb{S}_m^{-1}$ and $\mathbb{G}^m$.

2.1 Discretisation of Model Problem

Recall that the model problem in Definition 1.1.2 is posed on an infinite domain. In order to solve it numerically we must truncate it to a region of interest. As part of this truncation the Sommerfeld radiation condition must be approximated (1.6): we use the Perfectly Matched Layer (PML), see §1.3. The solution of the model problem can be approximated by solving the Helmholtz problem on the rectangle $[0, R]^2$ with PML boundary conditions approximating the Sommerfeld radiation condition on the sides. It is assumed that the support of f and $(1 - c)$ are contained within the rectangle inside the PML variables. The problem can always be scaled so that $R = 1$ and so without loss of generality we choose $R = 1$. We define the PML on $[0, 1]^2$ in §2.1.1.

We outline two discretisations of the truncated model problem with PML, a finite difference discretisation §2.1.2 and a finite element discretisation §2.1.3, both used extensively in numerical experiments later. The matrices A and A_{abs} that we create using these discretisations are specific cases of the discretisation matrices A and A_{abs} defined in Definitions 1.9.2 and 1.9.3 (recall that A_{abs} de-

notes the discretisation matrix where absorption is added to the wavenumber, i.e. $k(x) = k_R(x) + ik_I(x)$).

2.1.1 PML

We use the version of Perfectly Matched Layers developed by Collino and Tsogka [20]. For the model problem in Definition 1.1.2 we use the PML on all four

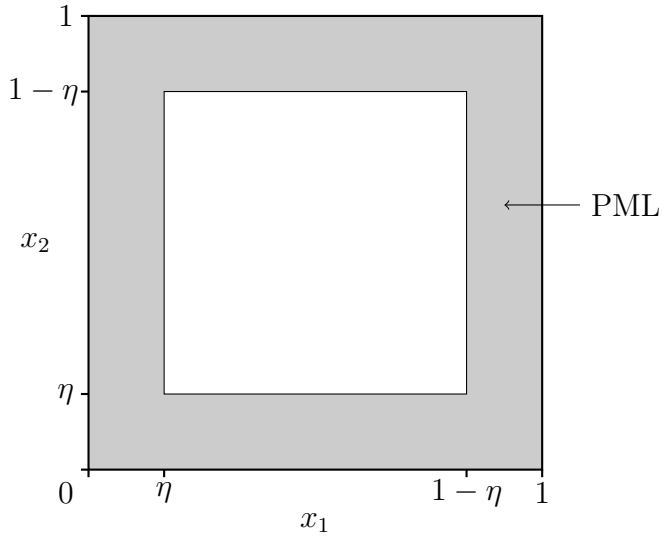


Figure 2-1: Sketch of PML

sides (see Figure 2-1). To implement the PMLs, the differential operators are transformed as follows

$$\begin{aligned} \frac{\partial}{\partial x_1} \quad &\text{is replaced by} \quad \theta_1(x_1)\frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \quad &\text{is replaced by} \quad \theta_2(x_2)\frac{\partial}{\partial x_2} \end{aligned}$$

where

$$\theta_1(x) = \theta_2(x) := \frac{1}{1 + \mathrm{i}\frac{\phi(x)}{k}} \tag{2.1}$$

and $\phi(x)$ is a cut-off function which has the value C/η at the boundary and 0 away from the boundary and is given by

$$\phi(x) := \begin{cases} \frac{C}{\eta}\left(\frac{x-\eta}{\eta}\right)^2 & \text{if } 0 \le x \le \eta, \\ 0 & \text{if } \eta < x < 1-\eta, \\ \frac{C}{\eta}\left(\frac{x-1+\eta}{\eta}\right)^2 & \text{if } 1-\eta \le x \le 1, \end{cases} \tag{2.2}$$

see Figure 2-2.

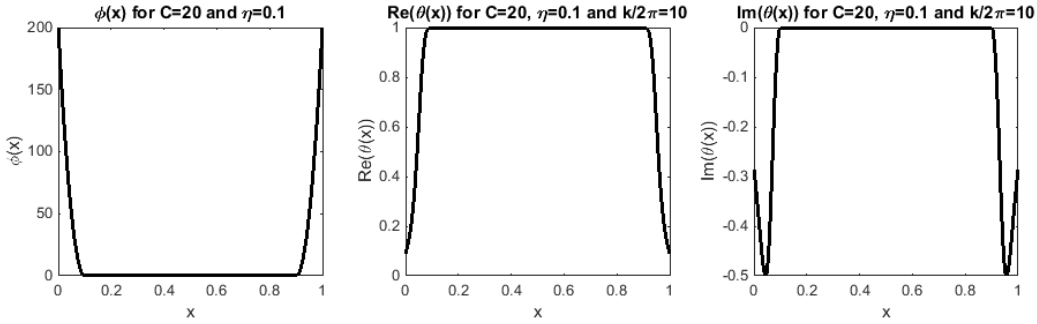


Figure 2-2: Graphs of the functions $\phi(x)$ and $\theta(x) := \theta_1(x) = \theta_2(x)$ used in the PML in Figure 2-1.

We assume that η, the width of the PML, is of the order of a wavelength and C is some positive constant independent of k. We can see from Figure 2-2 that the transformations only alter the differential operator in the PML region, as elsewhere ϕ is 0 and hence the θ_is are just 1. Recall from §1.3 that within the PML region the transformation causes exponential decay of incident waves [20] [69, §3.3.4], mimicking the Sommerfeld radiation condition and aiming to ensure that there are no reflections. We assume the support of f and c (i.e. the domain of interest) are within the $[\eta, 1-\eta]^2$ box.

Using these transformations, the Helmholtz equation becomes

$$\begin{aligned} \left(\theta_1(x_1)\frac{\partial}{\partial x_1}\theta_1(x_1)\frac{\partial}{\partial x_1} + \theta_2(x_2)\frac{\partial}{\partial x_2}\theta_2(x_2)\frac{\partial}{\partial x_2} + k^2\right)u(x) &= -f(x), \qquad x \in \Omega, \\ u(x) &= 0, \qquad x \in \partial\Omega, \end{aligned}$$

where $\Omega = [0,1]^2$. We divide through by $\theta_1(x_1)\theta_2(x_2)$ to give a symmetric form

of this PDE

$$\left(\frac{\partial}{\partial x_1}\frac{\theta_1(x_1)}{\theta_2(x_2)}\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2}\frac{\theta_2(x_2)}{\theta_1(x_1)}\frac{\partial}{\partial x_2} + \frac{k^2}{\theta_1(x_1)\theta_2(x_2)}\right) u(x) = -f(x), \qquad x \in \Omega, \qquad u(x) = 0, \qquad x \in \partial\Omega, \tag{2.3}$$

where $\Omega = [0,1]^2$ (observe the right hand side of (2.3) is still f, since f is assumed not to have support except where $\theta_1(x_1)$ and $\theta_2(x_2) = 1$).

We impose a homogeneous Dirichlet condition on the boundary of the $[0,1]^2$ box with this method, but this Dirichlet condition should be viewed as part of the PML, which is approximating the Sommerfeld radiation condition.

We describe the discretisation of this PDE using finite difference and finite element methods in the next two sections.

2.1.2 Finite Difference Discretisation

In practice most seismic problems in the frequency domain are solved using high-order finite-difference methods such as those outlined in one and two dimensions in [103]. For simplicity we here use only a double application of the central difference stencil for the first order derivatives, see left and middle of Figure 2-3, which normally reduces to the standard 5-point finite difference stencil for second order derivatives, see right of Figure 2-4. Slightly adapted versions of the preconditioners described below could also be applied to higher order discretisations.

We discretise the domain with a regular $(n+2) \times (n+2)$ grid, see Figure 2-4. Since the solution vanishes at the end of any row or column (due to the homogeneous Dirichlet boundary), there are n degrees of freedom per row. Since the boundary rows are given by the Dirichlet condition, the total number of degrees of freedom is $N = n^2$. The grid spacing is $h := 1/(n+1)$. Let $u_{i,j}$, $f_{i,j}$, $(\theta_1)_{i,j}$ and $(\theta_2)_{i,j}$ denote the corresponding functions evaluated at the points (ih, jh) (recall the definitions of PML variables $\theta_1(x_1)$ and $\theta_2(x_2)$ from (2.1)-(2.3)).

Then the finite difference approximation of (2.3), using the central difference

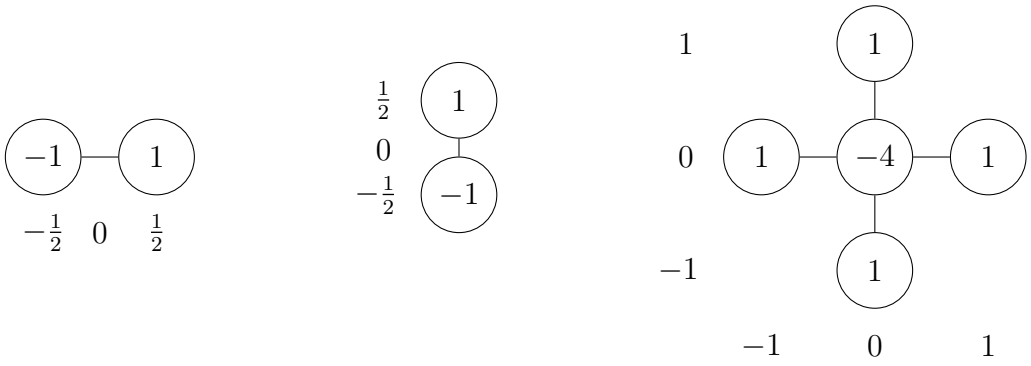


Figure 2-3: Left and middle: the central difference stencils for the first order derivatives, $\partial/\partial x_1$ and $\partial/\partial x_2$ respectively. Right: the 5-point finite difference stencil for second order derivatives $\partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$. Note the distance to the points in the central distance stencils are half that for the 5 point stencil.

stencils in Figure 2-4, is given by

$$\begin{aligned}
&\frac{1}{h^2}\left(\frac{(\theta_1)_{i-\frac{1}{2},j}}{(\theta_2)_{i-\frac{1}{2},j}}u_{i-1,j} + \frac{(\theta_1)_{i+\frac{1}{2},j}}{(\theta_2)_{i+\frac{1}{2},j}}u_{i+1,j} + \frac{(\theta_2)_{i,j-\frac{1}{2}}}{(\theta_1)_{i,j-\frac{1}{2}}}u_{i,j-1} + \frac{(\theta_2)_{i,j+\frac{1}{2}}}{(\theta_1)_{i,j+\frac{1}{2}}}u_{i,j+1}\right) \\
&+\left(\frac{k^2(x)}{(\theta_1)_{i,j}(\theta_2)_{i,j}} - \frac{1}{h^2}\left(\frac{(\theta_1)_{i-\frac{1}{2},j}}{(\theta_2)_{i-\frac{1}{2},j}} + \frac{(\theta_1)_{i+\frac{1}{2},j}}{(\theta_2)_{i+\frac{1}{2},j}} + \frac{(\theta_2)_{i,j-\frac{1}{2}}}{(\theta_1)_{i,j-\frac{1}{2}}} + \frac{(\theta_2)_{i,j+\frac{1}{2}}}{(\theta_1)_{i,j+\frac{1}{2}}}\right)\right)u_{i,j} \\
&= f_{i,j}, \qquad i,j \in \{1,\ldots,n\},
\end{aligned} \tag{2.4}$$

for any points on the edge of the domain, e.g. $u_{0,j}$, their value is 0. We briefly discuss how the derivation (2.4) was obtained, as the PML variables make it slightly non-standard, giving rise to half-value indices, e.g. $i - 1/2$ in the first term. Recall that the 5-point difference stencil arises from two applications of the central difference stencil for first order derivatives, see Figure 2-3. Looking at (2.3), it can be seen that for each term in the PDE, one application of the central difference stencil is performed before multiplication by $\frac{\theta_i}{\theta_j}$ where $i,j \in \{1,2\}$, $i \neq j$ and the other application of the central difference stencil is performed afterwards. Hence the θ_is are evaluated half-way between mesh entries.

We denote the linear system formed from (2.4) by $A\boldsymbol{u} = \boldsymbol{f}$ and similarly we denote the linear system (2.4) with $k(x) = k_R(x) + ik_I(x)$ by $A_{\text{abs}}\boldsymbol{u}_{\text{abs}} = \boldsymbol{f}$. The matrices A and A_{abs} are of size $N \times N$ with N^2 entries (where $N = n^2$). Assuming the nodes are ordered lexicographically (as shown in Figure 2-4) the

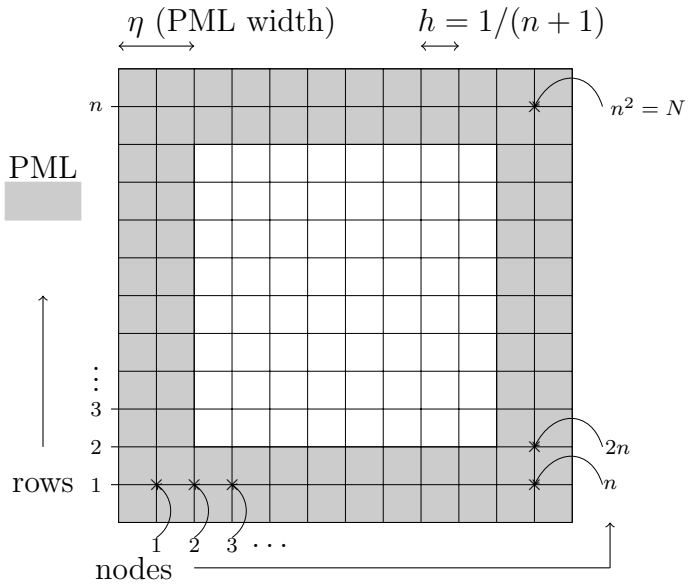


Figure 2-4: The discretisation grid with PMLs. Note the PML width is not to scale as it is generally more than 2 rows, for instance we use 12 rows in many of our experiments. Note that for an accurate solution h will be of the order of a wavelength or smaller.

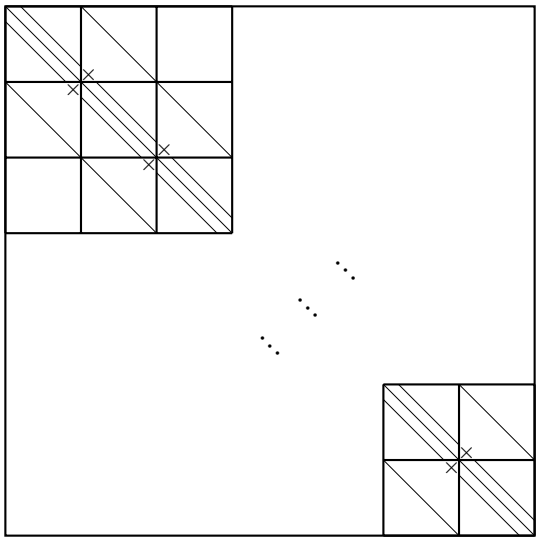


Figure 2-5: The block tridiagonal linear system, notice the lines at 0, ± 1 and $\pm n$ from the leading diagonal. Crosses indicate missing entries from stencils on boundary nodes.

matrices will also be block tridiagonal as in Figure 2-5. We use the finite difference discretisation in numerical experiments in §5.1.

2.1.3 Finite Element Discretisation

We have previously described the low-order finite element discretisation we shall use in this thesis, in §1.4. However, instead of constructing the finite element discretisation for the Interior Impedance Problem (as in §1.4), we construct it for the Helmholtz model problem with PML, see §2.1.1.

Remark 2.1.1. *We note that after multiplying* (2.4) *through by* h^2*, the resulting linear systems of low-order FEM or FD discretisations are nearly identical for interior nodes, if the integrals arising from the zero order term and the right-hand side in the FEM are evaluated using a quadrature rule that uses only nodal points (a different quadrature rule to that described previously in §1.4). This similarity is due to the fact that using this type of quadrature results in the mass matrix being diagonalised, due to a phenomena known as mass lumping [116, §16.2.4].*

2.1.4 Properties of A and A_{abs}

Note that the discretisation processes in §2.1.2 and §2.1.3 preserve certain properties of the underlying equation (2.3). In particular (2.3) is symmetric and complex valued and the matrices A and A_{abs} created using the FD or FE methods are also symmetric and complex valued. A crucial consequence is that the matrices A and A_{abs} are not Hermitian.

2.2 Outline of Sweeping Preconditioner

To understand the construction of the sweeping preconditioner to be applied to the discretisation matrix A, we first need to introduce a block decomposition of A.

Definition 2.2.1. ***(***M ***and*** Ω_m***)*** *The model problem in Definition 1.1.2 is discretised with PML using the methods in §2.1. The resulting truncated and discretised domain, is then divided up into* M *subdomains. The subdomains are* Ω_m*,* $m \in \{1, ..., M\}$*. Each subdomain contains at least one row of discretisation nodes*

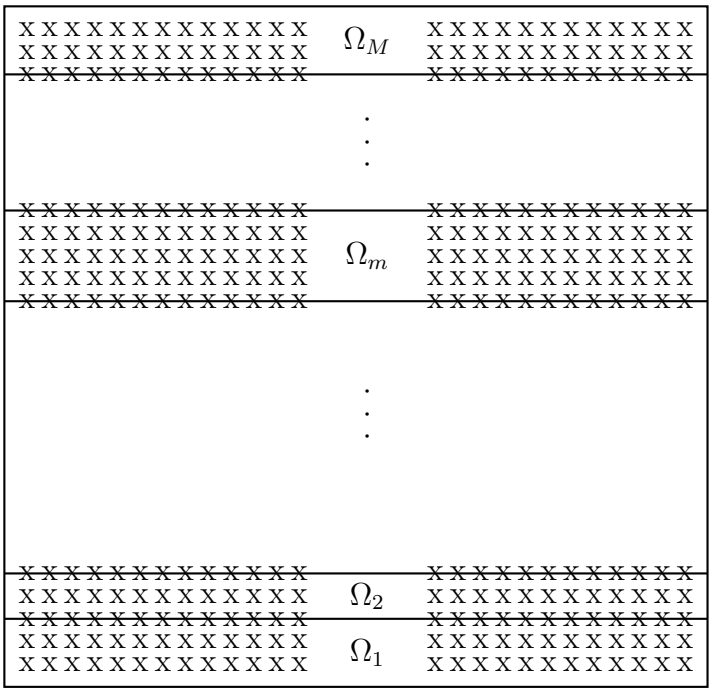


Figure 2-6: Subdomains Ω_m for $m \in \{1, ..., M\}$.

(see §2.1.2 and §2.1.3) and the horizontal boundaries of the subdomains lie along rows of discretisation nodes as in Figure 2-6.

We define a block structure for A, where blocks correspond to interactions between adjacent subdomains Ω_m for $m = \{1, \ldots, M\}$, or between any of these subdomains and itself. We assume the matrix A was created by using either the FD method in §2.1.2 or the lowest order FE method in §2.1.3. (Crucially these methods involve interactions only between nodes in the same and adjacent rows, so that the matrix A has a block-tridiagonal block decomposition, in the way we describe below. For higher order methods where interactions occur between more distant nodes, the block decomposition would require bigger blocks to include the additional non-zero entries of the matrix.)

We denote the block tridiagonal decomposition of the matrix A (where each block corresponding to the interaction between two adjacent subdomains or be-

tween the subdomain and itself), as follows:

$$A = \begin{pmatrix} \mathbb{A}_{1,1} & \mathbb{A}_{1,2} & & & 0 \\ \mathbb{A}_{2,1} & \mathbb{A}_{2,2} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \mathbb{A}_{M-1,M} \\ 0 & & & \mathbb{A}_{M,M-1} & \mathbb{A}_{M,M} \end{pmatrix}.$$

In particular, in this decomposition the nodes on boundaries of the domains Ω_m are assumed to be in the blocks corresponding to the lower domain. (Note that, due to how we have numbered the grid nodes, counterintuitively we go up the grid as we go down the matrix blocks, see Figure 2-4.)

Definition 2.2.2. ***(Number of Grid Rows*** D_m ***and*** D***)*** *For any* $m = \{1, \ldots, M\}$*,* D_m *is the number of grid rows contained within subdomain* Ω_m*. If all subdomains contain the same numbers of rows, we write* $D_m = D$*.*

Therefore, if Ω_m has D_m rows and Ω_j has D_j rows, then $\mathbb{A}_{m,j}$ is an $D_m n \times D_j n$ block. Since the subdomains may not be of equal size, the blocks may not be of equal size either.

We are interested in the following examples of subdivisions of A.

2.2.0.1 Example 1 $D_m = D = 1$

If $D_m = 1$ for every subdomain/for all $m = \{1, \ldots, M\}$ (i.e. each subdomain contains one row so that $D = 1$), then $\mathbb{A}_{m,j}$ is of size $n \times n$ and we write

$$A_{m,j} := \mathbb{A}_{m,j},$$

for each valid combination of m and j. In this case $M = n$.

The tridiagonal structure is then as follows:

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & & & 0 \\ A_{2,1} & A_{2,2} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & A_{n-1,n} \\ 0 & & & A_{n,n-1} & A_{n,n} \end{pmatrix}, \tag{2.5}$$

where each $A_{m,j}$ is an $n \times n$ block of A.

2.2.0.2 Example 2 $D = 1$, Finite Difference Discretisation

If we were to ignore the PML, by (2.4) the diagonal matrix blocks of the finite difference discretisation matrix would be

$$\mathbb{A}_{i,i} = A_{i,i} = \begin{pmatrix} k^2(x) - 4/h^2 & 1/h^2 & & & 0 \\ 1/h^2 & k^2(x) - 4/h^2 & \ddots & & \\ & & \ddots & \ddots & 1/h^2 \\ 0 & & & 1/h^2 & k^2(x) - 4/h^2 \end{pmatrix},$$
$$\text{for } i \in \{1, ..., M\}, \; x \in \Omega,$$

and the off-diagonal blocks would be

$$\mathbb{A}_{i,i+1} = A_{i,i+1} = A_{i+1,i} = \begin{pmatrix} 1/h^2 & & & 0 \\ & 1/h^2 & & \\ & & \ddots & \\ 0 & & & 1/h^2 \end{pmatrix} \text{ for } i \in \{1, ..., M-1\}.$$

Including the PMLs only changes the values of non-zero entries in these matrices corresponding to nodes near the boundary of Ω and within the PMLs depicted in grey in Figure 2-4 (right). Also, we can now see explicitly the matrix overall has the internal structure shown in Figure 2-5.

2.2.0.3 Example 3 $D_m > 1$ for at least one $m = \{1, \ldots, M\}$

In this case, each $\mathbb{A}_{i,j}$ can be built up from corresponding blocks of (2.5) and zero matrix blocks.

As an example, consider the case where Ω_1 includes the bottom three rows and Ω_2 the next two rows (see Figure 2-6). In this case we have:

$$\mathbb{A}_{1,1} = \begin{pmatrix} A_{1,1} & A_{1,2} & 0 \\ A_{2,1} & A_{2,2} & A_{2,3} \\ 0 & A_{3,2} & A_{3,3} \end{pmatrix} \tag{2.6}$$

$$\mathbb{A}_{1,2} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ A_{3,4} & 0 \end{pmatrix} \tag{2.7}$$

$$\mathbb{A}_{2,1} = \begin{pmatrix} 0 & 0 & A_{4,3} \\ 0 & 0 & 0 \end{pmatrix} \tag{2.8}$$

$$\mathbb{A}_{2,2} = \begin{pmatrix} A_{4,4} & A_{4,5} \\ A_{5,4} & A_{5,5} \end{pmatrix} \tag{2.9}$$

Note that the $\mathbb{A}_{m,m}$ may not be tridiagonal and the $\mathbb{A}_{m,m\pm1}$ may not be square.

Conditions 2.2.3. *From now on, for simplicity, we assume that $D_m = D \geq 1$ for all m. We also assume $D \ll n$ and D divides n so that $n = MD$.*

2.2.1 Definitions of $\mathbb{S}_m^{-1}$ and $\mathbb{G}^m$

Now we define some important notation and introduce some key ideas that are useful for formulating and understanding the sweeping preconditioner in the next two sections.

To form the basis of the preconditioner we apply block Gaussian elimination

to A to create a block LU decomposition.

$$A\boldsymbol{u} = \mathbf{f} = \begin{pmatrix} \mathbf{f}^1 \\ \mathbf{f}^2 \\ \vdots \\ \vdots \\ \mathbf{f}^M \end{pmatrix} = \mathbb{L}'\mathbb{U}'\boldsymbol{u} := \begin{pmatrix} \mathbb{S}_1 & & & & 0 \\ \mathbb{A}_{2,1} & \mathbb{S}_2 & & & \\ & \mathbb{A}_{3,2} & \ddots & & \\ & & \ddots & \ddots & \\ 0 & & & \mathbb{A}_{M,M-1} & \mathbb{S}_M \end{pmatrix} \times \dots \tag{2.10}$$

$$\dots \begin{pmatrix} \mathbb{I} & \mathbb{S}_1^{-1}\mathbb{A}_{1,2} & & & 0 \\ & \mathbb{I} & \mathbb{S}_2^{-1}\mathbb{A}_{2,3} & & \\ & & \ddots & \ddots & \\ & & & \ddots & \mathbb{S}_{M-1}^{-1}\mathbb{A}_{M-1,M} \\ 0 & & & & \mathbb{I} \end{pmatrix} \begin{pmatrix} \mathbf{u}^1 \\ \mathbf{u}^2 \\ \vdots \\ \vdots \\ \mathbf{u}^M \end{pmatrix}$$

where $\mathbf{u}^m$, $\mathbf{f}^m$ and $\mathbb{S}_m$ are as in the following definitions.

Definition 2.2.4. ***(Vectors $\mathbf{u}^m$, $\mathbf{f}^m$)*** *The vectors $\mathbf{u}^m$ and $\mathbf{f}^m$ are the solution vectors ($\mathbf{u}$ and $\mathbf{f}$ respectively, see §1.7.2) restricted to the nodes in Ω_m, respectively. (Recall that nodes on the boundaries of two subdomains are considered to be in the lower subdomain.)*

Definition 2.2.5. ***(Schur complement matrices $\mathbb{S}_m$)*** *The Schur complement matrices $\mathbb{S}_m$, are defined for $m \in \{1, \dots, M\}$ as*

$$\begin{aligned} \mathbb{S}_1 &:= \mathbb{A}_{1,1} \quad \textit{and} \\ \mathbb{S}_m &:= \mathbb{A}_{m,m} - \mathbb{A}_{m,m-1}\mathbb{S}_{m-1}^{-1}\mathbb{A}_{m-1,m}, \ \textit{for } m \in \{2, \dots, M\}. \end{aligned} \tag{2.11}$$

In §2.2.3 we see that the action of multiplying by $\mathbb{S}_m^{-1}$ is related to a sequence of Green's functions that correspond to a sequence of half-plane problems (recall Helmholtz half-plane problems in general from Definition 1.1.6). In particular, in Proposition 2.2.20 we show that the following sequence of Green's functions are the solution to the sequence of half-plane problems in Figure 2-7 and Definition 2.2.18.

Definition 2.2.6. ***(Green's functions G^m)*** *The Green's functions, G^m, where*

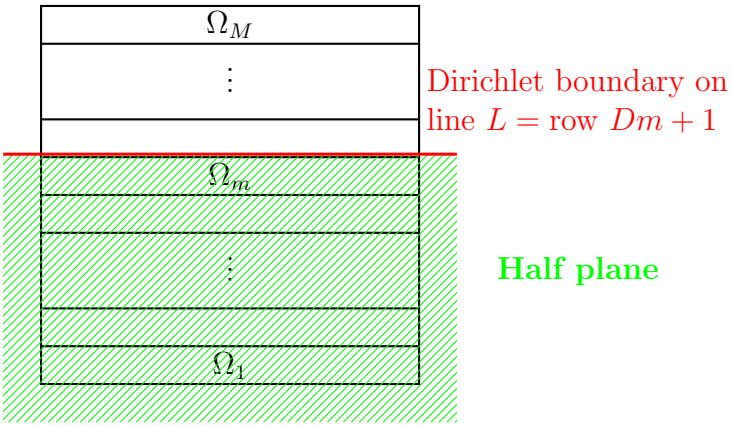


Figure 2-7: The half-planes associated with the half-plane problems in Definition 2.2.18. A Dirichlet condition is imposed on the half-plane boundary line L. The upper Dirichlet boundary line L is the row above Ω_m, i.e. row $Dm+1$.

$m \in \{1, \ldots, M\}$, *are*

$$G^m(\mathbf{x}, \mathbf{y}) := \frac{\mathrm{i}}{4} H_0(k\|\mathbf{x} - \mathbf{y}\|) - \frac{\mathrm{i}}{4} H_0(k\|\mathbf{x} - M(\mathbf{y})\|), \tag{2.12}$$

where $M(\mathbf{y})$ reflects the point $\mathbf{y}$ in the line L = row $Dm+1$ (see Figure 2-7), and H_0 is the Hankel function of the first kind, see §1.1.1.1.

In fact we later see that $\mathbb{S}_m^{-1} \approx \mathbb{G}^m$, where $\mathbb{G}^m$ are matrices with entries that are point evaluations of G^m, defined as follows.

Definition 2.2.7. ***(Matrices $\mathbb{G}^m$)*** *For each $m \in \{1, \ldots, M\}$, $\mathbb{G}^m$ is defined as*

$$(\mathbb{G}^m)_{i,j} := G^m(x_i, y_j),$$

where x_i, y_j are the nodes in Ω_m, ordered lexicographically from the bottom left, see example in Figure 2-8. (Recall that nodes on the boundaries of two subdomains are considered to be in the lower subdomain.)

Note especially the difference between G^m and $\mathbb{G}^m$, the former are the Green's functions in Definition 2.2.6 and the latter are the matrices formed by evaluating the Green's function G^m in Definition 2.2.7.

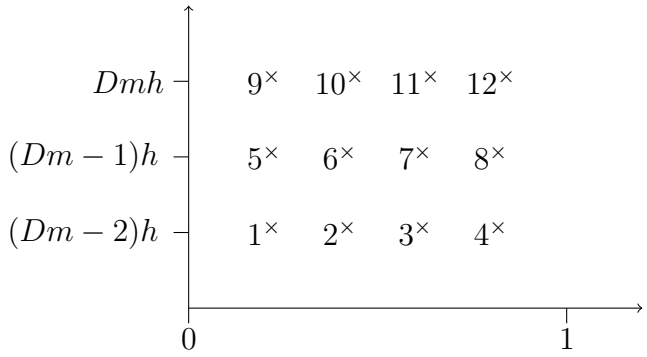


Figure 2-8: Assuming $D = 3$, $n = 4$, the nodes considered in $\mathbb{S}_m^{-1}$ and $\mathbb{G}^m$ are row Dm (for which the y-coordinate is Dmh) and the $D - 1 = 2$ rows below it; nodes have indices shown.

2.2.2 Discussion of First Key Idea: Sweeping as a block Thomas Algorithm

In §1.7.2 we saw the first key idea of sweeping preconditioners: sweeping through subdomains and creating cheap approximations to multiplying by Schur complement matrices to solve the Helmholtz problem. Here we describe sweeping in full detail, using a block version of the Thomas algorithm. Through describing the sweeping action we reveal the origins of the matrices $\mathbb{S}_m$ (see Definition 2.2.5) in the context of sweeping preconditioners. This information later allows us to establish a relationship between $\mathbb{S}_m^{-1}$ and $\mathbb{G}^m$ and explain how the relationship motivates approximating $\mathbb{S}_m^{-1}$ in 2.2.2.

The block structure, seen in Examples 1, 2, and 3 above, allows us to create a block decomposition of A or A_{abs} that forms the basis for the sweeping preconditioner. In particular, A and A_{abs} subdivided as in Examples 1, 2 or 3 are block tridiagonal, so that the block Thomas algorithm can be applied to them. (The element-wise Thomas algorithm is given in [108].) Shanks described the decomposition in terms of the Thomas algorithm where one line was 'swept' at each step, i.e., for the matrix subdivision when $D = 1$ (as in Examples 1 and 2). Here we adapt the algorithm to describe a version with $D > 1$, so that in the algorithm D lines are swept at each step of the algorithm. (This generalises the discussion in [35] (where multiple rows were used in each block) and [101] (where the process was described in terms of the Thomas algorithm).) The decomposition found by

the block Thomas algorithm is of this form:

$$A = \mathbb{L}\mathbb{D}\mathbb{L}^T, \tag{2.13}$$

where $\mathbb{D}$ is a block diagonal matrix and $\mathbb{L}$ is a block lower triangular matrix. The decomposition is of this form with $\mathbb{L}$ and $\mathbb{L}^T$ because A is symmetric. (This is a variation on the block decomposition (2.10) above. Note that the Schur complement matrices $\mathbb{S}_m$ that arise are the same for both decompositions.)

The first step is to form the decomposition

$$A = \mathbb{L}_1 \begin{pmatrix} \mathbb{S}_1 & 0 & 0 & & 0 \\ 0 & \mathbb{S}_2 & \mathbb{A}_{2,3} & & \\ 0 & \mathbb{A}_{3,2} & \mathbb{A}_{3,3} & \ddots & \\ & & \ddots & \ddots & \mathbb{A}_{M-1,M} \\ 0 & & & \mathbb{A}_{M,M-1} & \mathbb{A}_{M,M} \end{pmatrix} \mathbb{L}_1^t,$$

where $\mathbb{S}_1 = \mathbb{A}_{1,1}$, $\mathbb{S}_2 = \mathbb{A}_{2,2} - \mathbb{A}_{2,1}\mathbb{S}_1^{-1}\mathbb{A}_{1,2}$ and

$$\mathbb{L}_1 = \begin{pmatrix} \mathbb{I} & 0 & & 0 \\ \mathbb{A}_{2,1}\mathbb{S}_1^{-1} & \mathbb{I} & & \\ & & \ddots & \\ 0 & & & \mathbb{I} \end{pmatrix},$$

and where $\mathbb{I}$ denotes the $n \times n$ identity matrix. We begin with the first 2×2 square of blocks in the top left. In this first step of the sweep, the Schur complement matrices $\mathbb{S}_1$ and $\mathbb{S}_2$ on the diagonal have replaced the original four blocks of the matrix A and $\mathbb{L}_1$ is the first lower triangular matrix.

We now 'sweep through' the whole matrix A, creating Schur complements $\mathbb{S}_m$ and lower triangular matrices $\mathbb{L}_m$ for every 2×2 square of blocks in turn. In the end, we have $A = \mathbb{L}\mathbb{D}\mathbb{L}^T$ with

$$\mathbb{D} = \begin{pmatrix} \mathbb{S}_1 & & & 0 \\ & \mathbb{S}_2 & & \\ & & \ddots & \\ 0 & & & \mathbb{S}_M \end{pmatrix},$$

where the $\mathbb{S}_m$s are Schur complements as in Definition 2.2.5, and

$$\mathbb{L} := (\mathbb{L}_1, \dots, \mathbb{L}_{M-1}) = \begin{pmatrix} \mathbb{I} & 0 & & 0 \\ \mathbb{A}_{2,1}\mathbb{S}_1^{-1} & \mathbb{I} & & \\ & \ddots & \ddots & \\ 0 & & \mathbb{A}_{M,M-1}\mathbb{S}_{M-1}^{-1} & \mathbb{I} \end{pmatrix}.$$

At each step of the block Thomas algorithm we create the next Schur complement matrix $\mathbb{S}_m^{-1}$.

Note that creating this decomposition assumes that the matrices $\mathbb{S}_m$ are invertible. In the scalar (as opposed to the block) case, [50, Theorem 4.1.2, Corollary 4.2.3] proves that the decomposition exists for real-symmetric, positive-definite matrices (in the complex case the equivalent conditions require Hermitian positive-definite matrices). However, the system matrix is not positive definite in general; for example, for k sufficiently large the FEM discretisation is not coercive by [105, Lemma 6.2]. (In some instances it is known that A_{abs} is coercive, for example, when it satisfies the conditions in [53, Lemma 2.4]). However, in practice the sweeping preconditioners based upon the decomposition are effective in a variety of situations see [34, 35] and our experiments in §5.

Inverting $A = \mathbb{L}\mathbb{D}\mathbb{L}^T$ we have:

$$\begin{pmatrix} \mathbb{A}_{1,1} & \mathbb{A}_{1,2} & & 0 \\ \mathbb{A}_{2,1} & \mathbb{A}_{2,2} & \ddots & \\ & \ddots & \ddots & \mathbb{A}_{M-1,M} \\ 0 & & \mathbb{A}_{M,M-1} & \mathbb{A}_{M,M} \end{pmatrix}^{-1}$$

$$= \left(\mathbb{L}_1^t\right)^{-1} \dots \left(\mathbb{L}_{M-1}^t\right)^{-1} \begin{pmatrix} \mathbb{S}_1^{-1} & & & 0 \\ & \mathbb{S}_2^{-1} & & \\ & & \ddots & \\ 0 & & & \mathbb{S}_M^{-1} \end{pmatrix} \mathbb{L}_{M-1}^{-1} \dots \mathbb{L}_1^{-1}. \tag{2.14}$$

In its present form, (2.14) is not a computationally-practical method for solving $A\boldsymbol{u} = \boldsymbol{f}$ due to the prohibitively large cost of constructing and applying this decomposition (which is just a direct method for solving this system). The next two lemmas show that the computation of the matrices $\mathbb{S}_i$ (and simultaneously

their inverses) are the bottleneck of this computation. (Recall n is the number of degrees of freedom in each row of the grid and the total number of degrees of freedom is $N = n^2$.) We retain the D-dependence to illustrate the importance of $D \ll n$ to the costings.

Lemma 2.2.8. *Constructing all the matrices* $\mathbb{S}_m$ *and* $\mathbb{S}_m^{-1}$ *costs* $\mathcal{O}(D^2n^4) = \mathcal{O}(D^2N^2)$.

Proof. Recall that $\mathbb{S}_1 = \mathbb{A}_{1,1}$. Since $\mathbb{A}_{1,1}$ is of size $Dn \times Dn$ the cost to invert it is $\mathcal{O}(D^3n^3)$. We construct subsequent $\mathbb{S}_i$ matrices via (2.11); note that the cost of multiplying by the diagonal matrices $\mathbb{A}_{m,m-1}$ and $\mathbb{A}_{m-1,m}$ and the subtraction are less than $\mathcal{O}(D^3n^3)$. Therefore, the dominant cost of finding each matrix $\mathbb{S}_m$ via (2.11) is calculating the inverse of $\mathbb{S}_{m-1}$. Therefore the cost of constructing all the matrices $\mathbb{S}_m$ and $\mathbb{S}_m^{-1}$ is of the order of inverting all the $\mathbb{S}_m$ matrices. Since there are n/D of them, the overall cost is $\mathcal{O}(D^2n^4) = \mathcal{O}(D^2N^2)$. □

Lemma 2.2.9. *Assuming the matrices* $\mathbb{S}_m^{-1}$ *are known, the cost of multiplying a vector by* (2.14) *is* $\mathcal{O}(Dn^3) = \mathcal{O}(DN^{3/2})$.

Proof. Multiplying by (2.14) involves multiplying by $\mathbb{D}^{-1}$, which consists of n/D dense $\mathbb{S}_m^{-1}$ matrices of size $Dn \times Dn$ and therefore costs $\mathcal{O}(D^2n^3) = \mathcal{O}(D^2N^{3/2})$. Similarly, when multiplying by each $\mathbb{L}_m^{-1}$ and $(\mathbb{L}_m^T)^{-1}$ matrix, the main cost comes from multiplying by the dense $\mathbb{S}_m^{-1}$ matrices. To derive this costing, we observe that the $\mathbb{L}_m$ matrices are an example of block Gauss transformation matrix [50, p95]. (Recall that Gauss transformation matrices take the following form: the identity matrix, with one column containing arbitrary entries from the leading diagonal downwards). The inverse of a Gauss transform matrix is simply the original Gauss transform matrix with the non-zero entries below the diagonal multiplied by -1 [50, p97], so that

$$\mathbb{L}_m^{-1} = \begin{pmatrix} \mathbb{I} & & & & & 0 \\ & \ddots & & & & \\ & & \mathbb{I} & 0 & & \\ & & -\mathbb{A}_{m+1,m}\mathbb{S}_m^{-1} & \mathbb{I} & & \\ & & & & \ddots & \\ 0 & & & & & \mathbb{I} \end{pmatrix} \text{ for } m \in \{1, ..., M-1\}, \quad (2.15)$$

where the key entry is on the mth row. Multiplying by each $\mathbb{L}_m^{-1}$ costs $\mathcal{O}(D^2n^2)$ (as the $\mathbb{A}_{m+1,m}$ matrices are diagonal, this cost is solely due to the $\mathbb{S}_m^{-1}$ matrix) and there are $2(n/D-1)$ of them giving the overall cost of $\mathcal{O}(Dn^3) = \mathcal{O}(DN^{3/2})$. $\square$

The key idea of the sweeping preconditioner (as first discussed in §1.7.2) is to seek cheap approximation of the action of the $\mathbb{S}_m^{-1}$ matrices (possibly using cheap approximate inverses of $\mathbb{S}_m$ for $m \in \{1, ..., M-1\}$), since the bottleneck in using (2.14) to compute A^{-1} is multiplying by the $\mathbb{S}_i^{-1}$ matrices. Then we have a cheap approximation to the action of A^{-1} (2.14), which serves as the preconditioner. Note that to multiply by (2.14) the matrices $\mathbb{S}_i^{-1}$ do not need to be constructed explicitly, an approximation of their action is sufficient.

To make 'cheaply' precise, if the cost is the ideal of $\mathcal{O}(Dn)$ for multiplying by each $\mathbb{S}_i^{-1}$, the overall cost of multiplying by the approximation to A^{-1} (i.e. the preconditioner) is $\mathcal{O}(N)$.

Remark 2.2.10. *The block structure/subdivision and hence the preconditioner can equally be applied to the matrix A_{abs}, with the same complexity analysis described above. (Note A_{abs} is the discretisation of the same PDE, using the same FD or FE methods, the only difference being the value of k.)*

In fact the focus of this thesis is to investigate the effect of including absorption in the preconditioner (i.e. constructing the preconditioner from the matrix A_{abs} rather than A). Recall that adding some absorption can be beneficial to the preconditioner see §1.9.2.3.

2.2.3 Discussion of the Second Key Idea: Connection Between $\mathbb{S}_m^{-1}$ and $\mathbb{G}^m$

Recall from §1.7.2, the second key idea is to cheaply approximate the action of the $\mathbb{S}_m^{-1}$ matrices, is motivated by the fact that

the action of $S^{-1} \approx$ the action of a Helmholtz solution operator that can be shown to be low-rank.

We establish this fact in two key stages:

1. $\mathbb{S}_m^{-1} \approx \mathbb{G}^m$ (recall Definitions 2.2.5 and 2.2.7);

2. each G^m admits a separable expansion for disjoint subdomains of Ω_m (which we call weakly admissible, something formalised later in §4.2.2).

Putting 1. and 2. together we see that admissible off-diagonal blocks of $\mathbb{S}_m^{-1}$ have low-rank properties (i.e. that they admit good quality, low-rank approximations).

2.2.3.1 Stage 1: $\mathbb{S}_m^{-1} \approx \mathbb{G}^m$

Claim 2.2.11. $\mathbb{S}_m^{-1} \approx \mathbb{G}^m$.

The connection between $\mathbb{S}_m^{-1}$ and $\mathbb{G}^m$ in Claim 2.2.11 was first made by Engquist and Ying in [34], though only explicitly stated for the case $D = 1$. For the rest of this section we work through the argument that justifies Claim 2.2.11, based on the arguments in [34], adapting it to allow for $D \geq 1$. The argument is heuristic, but the numerical methods (which are based on theory that relies upon this claim) work well in practice.

In order to prove Claim 2.2.11, we show that multiplication by $\mathbb{S}_m^{-1}$ is a discrete version of an integral operator involving the Green's functions G^m in Definition 2.2.6, see Lemma 2.2.12.

Lemma 2.2.12. *Multiplication by $\mathbb{S}_m^{-1}$ is a discrete version of the operator*

$$g(x) \to \int_{\Omega_m} G^m(x,y)g(y)dy\Big|_{x\in\Omega_m}, \tag{2.16}$$

where $g(x) \in L^2(\Omega_m)$ is an arbitrary function.

In order to prove Lemma 2.2.12, we first prove several intermediate results.

We find that the Green's functions for the half-plane problems in Figure 2-7 are G^m (see Proposition 2.2.20), hence the presence of G^m in Lemma 2.2.12. Then we recall that each entry of each $\mathbb{G}^m$ is formed by evaluating each G^m on a pairs of nodes in Ω_m. Each entry of each $\mathbb{S}_m$ is also formed by looking at the interaction between a pair of nodes coming from either the FE or FD method, Claim 2.2.11 then follows.

To prove Lemma 2.2.12 and Claim 2.2.11, firstly we isolate the mth Schur complement matrix $\mathbb{S}_m^{-1}$.

Definition 2.2.13. ***(Matrix*** A_m^{-1}***)*** *We truncate the matrix decomposition (2.14) to the* m*th row or layer of blocks to obtain:*

$$A_m^{-1} := \begin{pmatrix} \mathbb{A}_{1,1} & \mathbb{A}_{1,2} & 0 & 0 \\ \mathbb{A}_{2,1} & \mathbb{A}_{2,2} & \ddots & 0 \\ 0 & \ddots & \ddots & \mathbb{A}_{m-1,m} \\ 0 & 0 & \mathbb{A}_{m,m-1} & \mathbb{A}_{m,m} \end{pmatrix}^{-1} =$$

$$\left(\mathbb{L}_1^t\right)^{-1} \dots \left(\mathbb{L}_{m-1}^t\right)^{-1} \begin{pmatrix} \mathbb{S}_1^{-1} & 0 & 0 & 0 \\ 0 & \mathbb{S}_2^{-1} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbb{S}_m^{-1} \end{pmatrix} \mathbb{L}_{m-1}^{-1} \dots \mathbb{L}_1^{-1}. \tag{2.17}$$

Proposition 2.2.14. *When the decomposition of* A_m^{-1} (2.17) *is multiplied out, the* (m,m)*th block of* $A_m^{-1} = \mathbb{S}_m^{-1}$*.*

Proof. Obtained by elementary linear algebra. □

The matrix A_m^{-1} is a discretisation matrix of a different Helmholtz problem than in Definition 1.1.2, next we investigate what this different problem is and find that it is the half-plane problem in Figure 2-7, with the associated Green's function G^m in Definition 2.2.6 and in Lemma 2.2.12.

Definition 2.2.15. ***(Domain*** $\widehat{\Lambda}_m$***)*** *Let the domain*

$$\widehat{\Lambda}_m := (0,1) \times (0, (Dm+1)h). \tag{2.18}$$

$\widehat{\Lambda}_m$ is the rectangular domain that is of interest to our half-plane problems. Note that A_m^{-1} has PMLs on three sides of $\widehat{\Lambda}_m$ (namely the south, east and west sides). Therefore, since the PMLs approximate the Sommerfeld Radiation Condition, A_m^{-1} can be thought of as a discretisation of a Green's function corresponding to a half-plane problem (recall Definition 1.1.6), with domain as follows.

Definition 2.2.16. ***(Half-plane Domain*** Λ_m***)*** *Let the half-plane domain*

$$\Lambda_m := (-\infty, +\infty) \times (-\infty, (Dm+1)h]. \tag{2.19}$$

The boundary condition at the top of the half-plane must be a Dirichlet condition. When we truncated A_m^{-1}, we removed the entries corresponding to the $Dm+1$th row upwards, setting the entries in the value of the solution on the $Dm+1$th row to zero or equivalently imposing a Dirichlet condition.

We now have enough information to state the half-plane problems that the matrix A_m^{-1} can be considered a discretisation of.

Following a similar process to §1.7.2, we consider a source function restricted to having non-zero entries on only Ω_m as follows.

Definition 2.2.17. ***(Restricted Source $\widehat{f}^m$)*** *Let $\widehat{f}^m(x)$ be defined for $x \in \Lambda_m$ as*

$$\widehat{f}^m(x) := \begin{cases} f(x), & x \in \Omega_m \backslash ((0,1) \times [(Dm-D)h, (Dm-D+1)h)), \\ 0, & else \end{cases}$$

where f is the source function as in the model problem we wish to solve in Definition 1.1.2.

Definition 2.2.18. ***(Sequence of Half-Plane Problems with solution $\widehat{u}^m$ (discretised $\widehat{\mathbf{u}}^m$) and source $\widehat{f}^m$ (discretised $\widehat{\mathbf{f}}^m$))*** *Let $\widehat{f}^m$ be defined as in Definition 2.2.17. Let $\widehat{u}^m$ be the solution of the Helmholtz equation*

$$\Delta_x \widehat{u}^m(x) + k^2 \widehat{u}^m(x) = -\widehat{f}^m(x), \quad x \in \Lambda_m, \tag{2.20}$$

with $k := \omega/c$, satisfying

$$\widehat{u}^m(x) = 0, \text{ for all } x \in \partial\Lambda_m = (-\infty, \infty) \times [(Dm+1)h], \tag{2.21}$$

and the Sommerfeld radiation condition (SRC)

$$\frac{x}{\|x\|} \cdot \nabla \widehat{u}^m(x) - \mathrm{i}k\widehat{u}^m(x) = o\left(\frac{1}{\|x\|}\right) \text{ as } \|x\| \to \infty. \tag{2.22}$$

Let the numerical solution of this adapted half-plane problem be given by the following system of linear equations:

$$\widehat{\mathbf{u}}^m = A_m^{-1} \widehat{\mathbf{f}}^m, \tag{2.23}$$

where $\widehat{\mathbf{u}}^m$ *and* $\widehat{\mathbf{f}}^m$ *are the solution and source of the discretised half-plane problems, i.e., approximations of* $\widehat{u}^m$ *and* $\widehat{f}^m$*, respectively.*

We have now identified the half-plane problems that A_m^{-1} are discretisations of. The domains and zero-Dirichlet conditions of the half-plane problems are as in Figure 2-7.

Definition 2.2.19. ***(The Half-Plane Green's Functions Associated with the Boundary-Value Problems Satisfied by*** $\widehat{u}^m$***)*** *The Green's functions for the half-plane problems in Definition 2.2.18 satisfy*

$$(\Delta_x + k^2)G^m(x,y) = -\delta(y-x) \text{ in the distributional sense for all } x, y \in \Lambda_m, \tag{2.24}$$

$$G^m(x,y) = 0, \text{ for all } x, y \in \partial\Lambda_m = (-\infty, \infty) \times [(Dm+1)h], \tag{2.25}$$

and

$$\frac{x}{\|x\|} \cdot \nabla_x G^m(x,y) - \mathrm{i}kG^m(x,y) = o\left(\frac{1}{\|x\|}\right) \text{ as } \|x\| \to \infty \text{ in the lower half-plane.} \tag{2.26}$$

Proposition 2.2.20. *The Green's functions for the half-plane problems in Definition 2.2.18, are the functions* G^m *the sums of Hankel functions in Definition 2.2.6.*

Proof. To prove this proposition it is sufficient to show that the Green's functions G^m satisfy Definition 2.2.19. Recall that the Hankel function is the fundamental solution (or free-space Green's function) to the Helmholtz equation and is therefore a solution of the full-plane Helmholtz problem with Sommerfeld Radiation condition (i.e. the model problem with $k(x) \equiv 1$) (see §1.1.1.1 and Definition 1.1.2). Therefore G^m clearly satisfies (2.24) and (2.26). The operator $M(y)$ in the definition of G^m (see Definition 2.2.6) reflects y in the boundary row $Dm+1$, so that

$$\begin{aligned} G^m(\mathbf{x},\mathbf{y}) =& \frac{\mathrm{i}}{4}H_0(k\|\mathbf{x}-\mathbf{y}\|) - \frac{\mathrm{i}}{4}H_0(k\|\mathbf{x} - M(\mathbf{y})\|) \\ =& \frac{\mathrm{i}}{4}H_0(k\|\mathbf{x}-\mathbf{y}\|) - \frac{\mathrm{i}}{4}H_0(k\|\mathbf{x} - \mathbf{y}\|) = 0 \text{ on } \partial\Lambda_m. \end{aligned} \tag{2.27}$$

Thus, by what is called the method of images, G^m vanishes on the Dirichlet boundary $L = \partial\Lambda_m$, as required by (2.25). □

Now we have all the definitions and results to prove Lemma 2.2.12 that we stated earlier.

Proof of Lemma 2.2.12. In Proposition 2.2.20, we established that G^m are the sequence of Green's functions corresponding to the sequence of half-plane problems in Definition 2.2.18 and Figure 2-7. Let $\widehat{\mathbf{u}}^m|_{\Omega_m}$ and $\widehat{\mathbf{f}}^m|_{\Omega_m}$ be the restrictions of $\widehat{\mathbf{u}}^m$ and $\widehat{\mathbf{f}}^m$ to Ω_m, respectively (where the nodes on the boundaries of two subdomains are considered to be part of the lower subdomain). In Proposition 2.2.14 we established that the (m, m)th block of A_m^{-1} is $\mathbb{S}_m^{-1}$. So the last line of (2.23) when multiplied out is

$$\widehat{\mathbf{u}}^m|_{\Omega_m} = \mathbb{S}_m^{-1}\widehat{\mathbf{f}}^m|_{\Omega_m}. \tag{2.28}$$

Therefore multiplication by $\mathbb{S}_m^{-1}$ corresponds to taking the source functions restricted to Ω_m, to the solutions restricted to Ω_m via the action of the Green's function operators restricted to Ω_m, proving the lemma statement. □

Key idea

We wish to particularly highlight here the key idea seen at the start of the chapter: "the action of a Schur complement matrix $\mathbb{S}^{-1} \approx$ action of a Helmholtz solution operator that can be shown to be low-rank". The following was established as part of the above proof:

$$\widehat{\mathbf{u}}^m|_{\Omega_m} = \mathbb{S}_m^{-1}\widehat{\mathbf{f}}^m|_{\Omega_m}, \tag{2.29}$$

where $\widehat{\mathbf{u}}^m|_{\Omega_m}$ and $\widehat{\mathbf{f}}^m|_{\Omega_m}$ are the restriction to Ω_m of the source and solution of the half-plane problems in Definition 2.2.18 respectively (where the nodes on the boundaries of two subdomains are considered to be part of the lower subdomain). Therefore the action of $\mathbb{S}_m^{-1}$ corresponds to partially solving a half-plane Helmholtz problem, taking a source function restricted to Ω_m, to observing the solution restricted to Ω_m, via the action of the Green's function operators restricted to Ω_m.

Remark 2.2.21. *The approximation $\mathbb{S}_m^{-1} \approx \mathbb{G}^m$ implies that $\mathbb{S}_m^{-1}$ should inherit properties of the Green's functions G^m used to define the $\mathbb{G}^m$ matrices. In particular, this allows us to obtain low-rank properties of $\mathbb{S}_m^{-1}$ (see next section).*

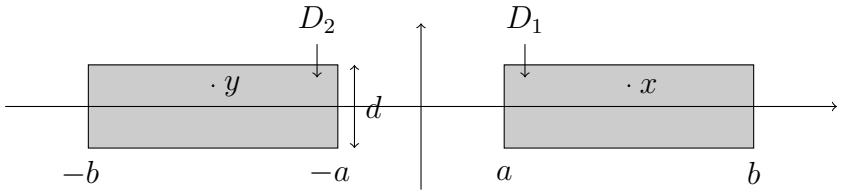


Figure 2-9: The domains of Theorem 2.2.22 (as in [34, Fig 2.2]).

We note that the analysis given ignores several sources of approximation error:

1. error from the PML approximation of the Sommerfeld Radiation Condition;
2. discretisation method error (reduces as $h \to \infty$);
3. in practice the method is used for inhomogeneous wavespeed models, so G^m is not the right Green's function for these problems and the actual Green's function for these problems in general isn't known explicitly.

2.2.3.2 Stage 2: the Green's Functions G^m Admit a Separable Expansions

Next we show that $G^m(x, y)$ admit low-rank separable expansions when x and y lie in separated domains. The outline for the low-rank results here is based on Engquist and Ying [34], where they make reference to earlier work by Rokhlin and Martinsson [82].

To begin we look at the following theorem due to Rokhlin and Martinsson [82] that shows the existence of a good low-rank separable expansion (see Definition 1.8.1) for the Hankel function.

Theorem 2.2.22. *([34, Theorem 2.4] due to Rokhlin and Martinsson [82, Theorem 2], wording and notation adapted) Let $a > 0$, $b > 0$, $d > 0$, $D_1 := [a, b] \times [-d/2, d/2]$, $D_2 := [-b, -a] \times [-d/2, d/2]$ and $x \in D_1, y \in D_2$, as in Figure 2-9. Given $\varepsilon \in (0, 1/2)$, there exists a constant $p \leqslant \log(2kb)|\log \varepsilon|^2$, a constant $C(d)$ such that $ka > C(d)|\log \varepsilon|$ and suitable functions $\{\phi_j, \chi_j\}_{j=1}^p$ such that*

$$\left| H_0(k\|x - y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y) \right| \leqslant \varepsilon, \tag{2.30}$$

holds for all $x \in D_1, y \in D_2$.

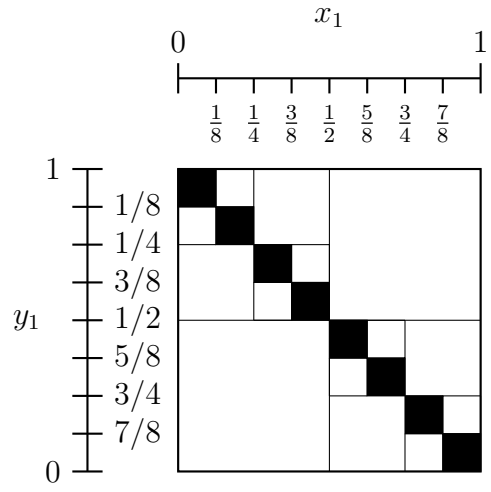


Figure 2-10: The admissible block structure of a matrix $\mathbb{G}^m$, admissible blocks in white. The axes on the side of the matrix show which parts of the matrix correspond to which values of x and y in the arguments of $G^m(x, y)$.

The separable expansion (3.33) is low-rank because the rank p is small; in this case the rank grows only weakly with k. Note that D_1 and D_2 (the domains of x and y respectively) are disjoint in Figure 2-9; this disjointness is necessary to avoid the singularity of the Hankel function at zero.

Since the Green's functions G^m are the sum of two Hankel functions (see (2.12)), the existence of a low-rank separable expansion for the Hankel function allows us to show the existence of low-rank separable expansions for each G^m and hence for $\mathbb{G}^m$. In particular we now consider finding low-rank matrix approximations for off-diagonal blocks of each $\mathbb{G}^m$ using Theorem 2.2.22.

Existing theory from [34] considers low-rank matrix approximations for off-diagonal blocks of $\mathbb{G}^m$ in the case $D = 1$. In this case, each $\mathbb{G}^m$ consists of point-evaluations $G^m(x, y)$ where x and y lie only on the line $x_2 = Dmh = mh$ (see Definition 2.2.7). Recalling the definition of $G^m(x, y)$ as the sum of two Hankel functions in (2.12), see Definition 2.2.6, it is sufficient for us to consider the Hankel function, with x and y in the box $B = [0, 1] \times [mh, (m + 2)h]$ (to see this, note that in the argument of second Hankel function in (2.12), $M(y)$ lives on the line $x_2 = (Dm + 2)h = (m + 2)h$).

We now recall Engquist and Ying's theorem, where they use the separable expansions for the Hankel function on a pair of domains D_1 and D_2 from Theorem

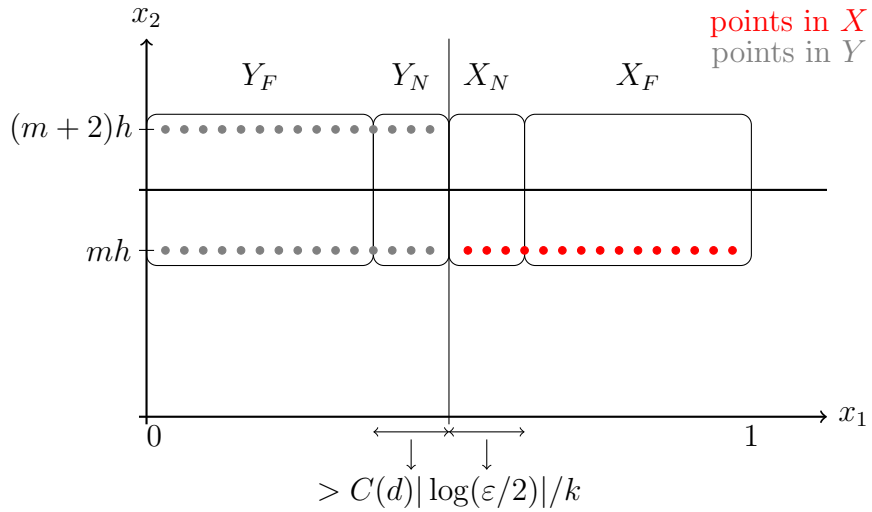


Figure 2-11: Based on [34, Fig 2.2]. Points in the sets X and Y from Theorem 2.2.23 that can be covered by domains X_F and Y_F, which satisfy the conditions of the domains D_1 and D_2 from Theorem 2.2.22 (note especially that the separation of X_F and Y_F satisfies the condition $ka > C(d)|\log(\varepsilon/2)|$).

2.2.22, where $D_1, D_2 \subset B = [0,1] \times [mh, (m+2)h]$. They obtain a low-rank approximation to the the upper right-hand block of $\mathbb{G}^m$ in Figure 2-10.

Theorem 2.2.23. *([34, Theorem 2.3] wording and notation adapted, note that we have added the condition in the square bracket to the statement of this result, we infer it is needed from their proof of this result.) Let $d = 2h$ and*

$$Y = \left\{(ih, mh), i = 1, \ldots, \frac{n}{2}\right\},$$
$$X = \left\{(ih, mh), i = \frac{n}{2} + 1, \ldots, n\right\}.$$

Given $\varepsilon > 0$, let

$$\begin{aligned} X_N &:= \left\{(p_1, p_2) \in X,\ p_1 \leq \frac{1}{2} + C(d)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\frac{1}{k}\right\}, \\ X_F &:= \left\{(p_1, p_2) \in X,\ p_1 > \frac{1}{2} + C(d)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\frac{1}{k}\right\}, \end{aligned} \tag{2.31}$$

$$
\begin{aligned}
Y_N &:= \left\{ (p_1, p_2) \in Y,\ p_1 \geq \frac{1}{2} - C(d) \left| \log\left(\frac{\varepsilon}{2}\right) \right| \frac{1}{k} \right\}, \\
Y_F &:= \left\{ (p_1, p_2) \in Y,\ p_1 < \frac{1}{2} - C(d) \left| \log\left(\frac{\varepsilon}{2}\right) \right| \frac{1}{k} \right\},
\end{aligned}
\tag{2.32}
$$

so that they form a partition of X and Y as in Figure 2-11, i.e.,

$$
\begin{aligned}
Y &= Y_N \cup Y_F, \\
X &= X_N \cup X_F.
\end{aligned}
$$

Let G^m be the (continuous) half-plane Green's functions of the Helmholtz operator for the domain $(-\infty, \infty) \times (-\infty, (m+1)h)$ with zero boundary condition. [h is chosen so that $hk^{-1} \sim 1$]. Then the matrix $(G^m(x,y))_{x \in X, y \in Y}$ is numerically low-rank. More precisely, for any $\varepsilon > 0$, there exists a constant $p = \mathcal{O}(\log k \left|\log \varepsilon\right|^2)$ and functions $\{\alpha_j(x)\}_{1 \leq j \leq p}$ for $x \in X$ and functions $\{\beta_j(y)\}_{1 \leq j \leq p}$ for $y \in Y$ such that

$$
\left| G^m(x,y) - \sum_{j=1}^{p} \alpha_j(x) \beta_j(y) \right| \leq \varepsilon, \quad \text{for } m \in \{1, ..., M\},\ x \in X,\ y \in Y.
$$

We give a full version of the proof of Theorem 2.2.23 in §4.2.6. Here we briefly state the idea of the proof. We note that Theorem 2.2.22 is valid for $x \in X_F$, and $y \in Y_F$, with X_F and Y_F depicted in Figure 2-11. (The theorem is valid on these domains because $\|x - y\|$ – in the argument of the Hankel function in (3.33) – is invariant under identical translations of $x \in D_1$ and $y \in D_2$. Therefore Theorem 2.2.22 is valid for all pairs of domains of the same size and relative positions as D_1 and D_2). Since the Green's functions G^m are the sum of two Hankel functions (see (2.12)), the existence of a separable expansion for each G^m on X_F and Y_F is obtained by adding together two separable expansions of the Hankel function. X_F and Y_F mostly cover the sets of points X and Y, see Figure 2-11. Therefore there exists a numerically low-rank expansion for the corresponding parts of the matrices $(G^m(x,y))_{x \in X, y \in Y}$ (the large bottom right-hand block of the matrix in Figure 2-12). The remaining points (located in X_N and Y_N in Figure 2-11) are sufficiently few (they are bounded above by $C(2h) \left| \log\left(\frac{\varepsilon}{2}\right) \right|$, for details see full proof in §4.2.6) that storing the rest of the matrix in a full-rank way does not affect the order of the size of the rank to

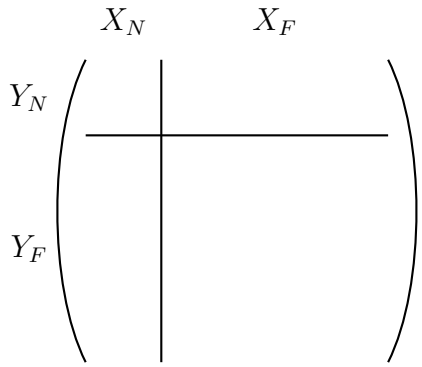


Figure 2-12: $(G^m(x,y))_{x\in X, y\in Y}$ split according to X_N, X_F, Y_N and Y_F.

approximate the whole matrix, giving the result.

Later we revisit the concepts of admissibility and admissible block structures like Figure 2-10, they originate in the study of $\mathcal{H}$-matrices, see §1.8.3 and §4.2.2.

Due to the fact that $\mathbb{S}_m^{-1} \approx \mathbb{G}^m$, the results about the low-rank approximation of each $\mathbb{G}^m$ imply that each $\mathbb{S}_m^{-1}$ should also admit good-quality, low-rank approximations on the admissible block structure. It is the $\mathcal{H}$-matrix framework (see §1.8.3 and §4.2.2-4.2.3.1), that provides a way to make use of this property of the $\mathbb{S}_m^{-1}$ matrices to approximate cheaply multiplication by the $\mathbb{S}_m^{-1}$ matrices (recall that cheaply approximate multiplication by the $\mathbb{S}_m^{-1}$ matrices was a key part of the sweeping preconditioner). Engquist and Ying approximate the $\mathbb{S}_m^{-1}$ matrices using the $\mathcal{H}$-matrix framework in their preconditioner in [34] and we perform variations on their experiments in §5.1.3-§5.1.4.

There are limitations to the scope of this low-rank theory for $\mathbb{G}^m$ and $\mathbb{S}_m^{-1}$, other than that it only considers the case $D = 1$. Another limitation of the theory is that the effect of adding absorption to the wavenumber k is not investigated. Also a constant number of grid points per wavelength is assumed, which for low-order methods does not hold in practice due to the need to combat the pollution effect, see §1.4.2. In this thesis we prove theorems that address these points.

2.3 Outline of the Following Chapters

Now we have described the sweeping preconditioner and expanded on the two key ideas, we give a summary of the rest of the thesis chapters.

In Chapter 3 we prove new results about the existence of a low-rank separable

expansions to the Hankel function; these can be viewed as new and extended versions of Theorem 2.2.22. Using these low-rank results for the Hankel function we then prove the existence of low-rank separable expansions for the Green's functions G^m. Crucially our theorems permit absorption to be added to the wavenumber, and this allows us to investigate the effect of absorption on the low-rank results. Also, we consider explicitly the k dependence of the sizes of the domains of the Hankel function, allowing us to consider domains with $D > 1$. We find that D can be relatively large for some blocks when absorption is included.

In Chapter 4 we prove new results about the existence of low-rank matrix approximations of off-diagonal blocks of $\mathbb{G}^m$ and thus their $\mathcal{H}$-matrix approximations, these can be viewed as new and extended versions of Theorem 2.2.23. We then perform numerical experiments, verifying these properties of $\mathbb{G}^m$. We also investigate the low-rank properties of $\mathbb{S}_m^{-1}$ (constructed using the FEM from §2.1.3), which we expect to behave similarly.

In Chapter 5 we perform numerical experiments where we construct preconditioners using the method described in this chapter, to investigate the effects of absorption on the iteration counts. The Schur complements for these experiments are approximated in different ways, specifically using Engquist and Ying's moving PML method [35] and their $\mathcal{H}$-matrix framework method [34].

Chapter 3

New Low-Rank Results for the Hankel and Green's Functions

3.1 Low-Rank Results

In this chapter we present our new low-rank results for the Hankel function. Recall that the Hankel function $H_0^{(1)}(k\|x-y\|)$ (1.3) is the fundamental solution to our model Helmholtz problem, see Definition 1.1.2 and §1.1.1.1.

We described the general importance of low-rank separable expansions to fundamental solutions in §1.8. However, in this thesis we are most concerned with the particular application of the low-rank separable expansion of the Helmholtz fundamental solution in 2D for sweeping preconditioners, seen in §2.2.3.

Recall from §2.2.3.1 that the key idea of the sweeping preconditioner was that the Schur Complement matrices $\mathbb{S}_m^{-1}$ (Definition 2.2.5) in (2.14) are approximately equal to $\mathbb{G}^m$ (Definition 2.2.7) a discretised version of the Green's function G^m (Definition 2.2.6) that is the sum of two Hankel functions. A low-rank result for the Hankel function by Rokhlin and Martinsson, Theorem 2.2.22, motivates a $\mathcal{H}$-matrix representation of the Schur Complement matrices $\mathbb{S}_m^{-1}$, as part of Engquist and Ying's Sweeping Preconditioner [34] for solution of the Helmholtz equation.

In §1.9 we also described that there are (under certain conditions) further benefits (i.e. further reduced GMRES iteration counts) if the preconditioner is constructed for the Helmholtz operator with absorption i.e. if a complex shift is

added to the wavenumber, replacing k with $k_R+\mathrm{i}k_I$. However, the low-rank theory behind Theorems 2.2.22 and 2.2.23 does not consider complex wavenumber.

Since the Rokhlin and Martinsson low-rank result Theorem 2.2.22 underlies the sweeping preconditioner, in order to analyse the sweeping preconditioner with complex wavenumbers, we need an analogue of the Rokhlin and Martinsson low-rank result that covers complex wavenumbers. We also need a low-rank result that allows us to do k-explicit analysis, including analysing the effect of domain sizes varying with k. This is because the fineness of the mesh, and hence the sets of points in the definition of $\mathbb{G}^m$, depend on k, e.g. $h \sim k^{-3/2}$ to counteract the pollution effect (see §1.4.2). This chapter is dedicated to obtaining such a result, first for the Hankel function H_0^1, then for the Green's function G^m.

We recall from §1.8.1 that it is important, when trying to obtain low-rank separable expansions to Helmholtz fundamental solutions (including the Hankel function in 2D), that the domains of the fundamental solution have both distance/separation and directionality imposed on them. To provide a further demonstration of this, we look at the Hankel function evaluated on two different pairs of domains as in Figure 3-1, where the top pair of domains is not separated and the bottom pair are separated. The Hankel function evaluated on these domains is displayed in Figure 3-2; the top of Figure 3-2 corresponds to the non-separated domains and the bottom to the separated domains. We see that, in the non-separated case (top) compared to the separated case (bottom), as well as the obvious presence of the singularity at $y = [0, 25]$, the amplitude of the oscillations is greater and the waves also turn through a greater angle from the x to the y-axes. The non-separated case is therefore inherently more 'complicated' and harder to approximate with a low-rank separable expansion.

3.2 Statement and Analysis of New Low-Rank Results for the Hankel Function

Definition 3.2.1. ***(Notation for domains and points)*** *Given* $1 > d > 0$ *and* $0 < a < b \leq 1$, *we define* $D_1 := [a, b] \times [0, d]$, $D_2 := [-b, -a] \times [0, d]$. *Let* $x = [x_1, x_2]^T \in D_1$ *and* $y = [y_1, y_2]^T \in D_2$ *(see Figure 3-3). Let*

$$\operatorname{diam}(D_1, D_2) := \max\{\operatorname{diam}(D_1), \operatorname{diam}(D_2)\} = \sqrt{(b-a)^2 + d^2}.$$

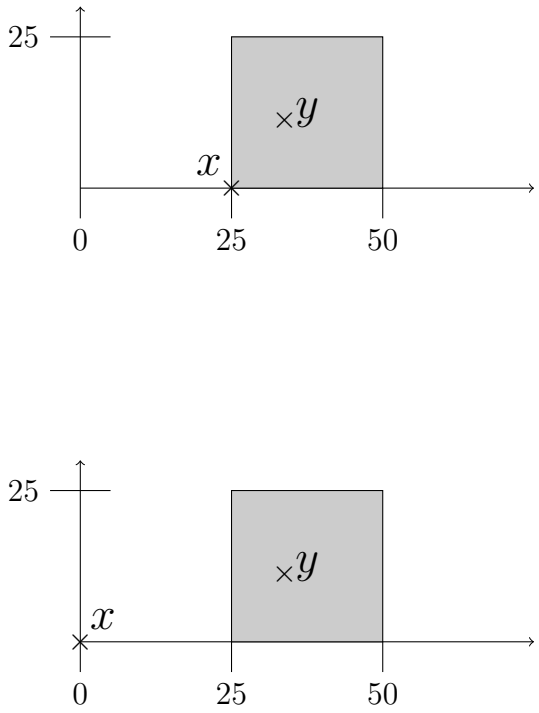


Figure 3-1: x and y values for Hankel function in Figure 3-2. Top: not separated, bottom: separated.

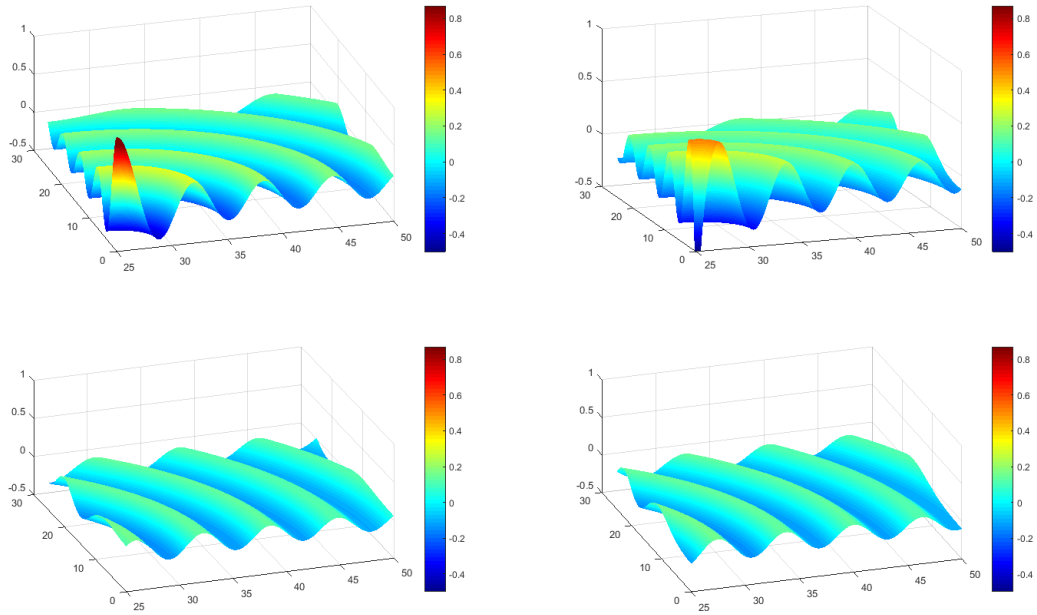


Figure 3-2: Left: real($H_0(k\|x-y\|)$). Right: imag($H_0(k\|x-y\|)$). Top: $x = [25, 0]$. Bottom $x = [0, 0]$. $y \in [25, 50] \times [0, 25]$, as in Figure 3-1.

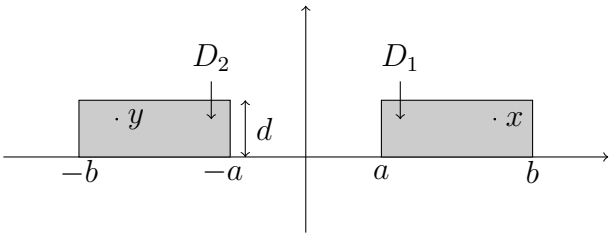


Figure 3-3: Domains of our new low-rank result.

We add absorption to the problem as follows:

Definition 3.2.2. ***(Absorption** k_I**)*** *The convention for adding absorption to the wavenumber is* $k := k_R + \mathrm{i}k_I$*, where* $k_R \geq 1$ *and* $0 \leq k_I \leq k_R$*.*

(For details on the different conventions used for including absorption, see §1.9.2.3. Note that in considering adding absorption in the range $0 \leq k_I \leq k_R$, we cover Shifted-Laplacian type shifts up to $\alpha \lesssim k^2$, which is the maximum shift considered in practice.)

We now state our new low-rank result with $k \in \mathbb{C}$.

Theorem 3.2.3. ***New Low-Rank Result for the Hankel Function*** *Let the domains and absorption be as defined in Definition 3.2.1 and Definition 3.2.2 respectively. Assume that for some constant* $\eta > 0$*,* $\eta \mathrm{dist}(D_1, D_2) > \mathrm{diam}(D_1, D_2)$*. Then there exists* $C_1 > 0$ *independent of* a*,* b*,* d*,* η *and* k*, such that, given* $\varepsilon \in (0, 1)$*, if*

$$\frac{k_R d^2}{a}\left[1 + \left|\log\left(4(k_R a)^2 + (k_R d)^2\right)\right|\right] \leq C_1 \varepsilon \exp(k_I a), \tag{3.1}$$

then there exist functions $\{\phi_j, \chi_j\}_{j=1}^{p}$*, where*

$$p = \left\lceil C_2(\eta) \max\left\{1, \log^2\left(C_3(\eta) \max\left\{\frac{\exp(-2k_I a)}{\varepsilon}, 1\right\}\right)\right\}\right\rceil, \tag{3.2}$$

and where C_2 *and* C_3 *depend only on* η*, such that*

$$\left|H_0(k\|x - y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y)\right| \leqslant \varepsilon, \tag{3.3}$$

for all $x \in D_1, y \in D_2$*.*

The proof of this result it contained in §3.4.

We highlight that Theorem 3.2.3 is proved in an entirely different way to how Rokhlin and Martinsson proved Theorem 2.2.22 and so the restrictive conditions on the domains and resulting low-rank separable expansion come in a slightly different form to theirs.

Having stated the result, there are several directions to take to investigate its implications.

1) The condition (3.1) is complicated and it is difficult to understand for which combinations of conditions on the domains and values of k and ε the theorem are valid. In §3.2.1 we specialise Theorem 3.2.3 into several different situations to better illustrate which combinations of conditions (3.1) allows.

   A slight variant of Theorem 3.2.3 is presented in §3.2.3; then we also specialise this variant into several different situations.

2) The expression for the rank p in (3.2) is quite complicated and we examine this further in §3.2.2.

3) We recall that as we increase absorption (i.e. as we increase k_I) the oscillations of the Hankel function are damped. Consequently, we expect it to be easier to approximate the Hankel function as we increase k_I. In §3.2.3 we give a summary of the ways in which we see improvements due to absorption and also present the variant of the new low-rank result for non-zero k_I.

4) It is natural to compare this result to Rokhlin and Martinsson's result (Theorem 2.2.22); we recall that the whole rationale behind proving Theorem 3.2.3 was to obtain an analogue of Rokhlin and Martinsson's result with complex k. The lemmas mentioned in 1) help us to do this comparison in §3.2.4.

   In the next four sub-sections we look at each of these points in order.

3.2.1 Domains for which Theorem 3.2.3 is valid

The condition (3.1) is complicated and we now specialise Theorem 3.2.3 into three situations, to illustrate for which combinations of conditions on the domains and values k and ε the theorem is valid.

The ideal situation would be one with

- large k_R values so that we can consider high-frequency Helmholtz problems.
- large domains, or equivalently large values of $b-a$ and d.
- small separation a between the domains.
- small rank p of the low-rank approximation.

However, due to the oscillatory nature of the Helmholtz fundamental solution, it is not to be expected that this ideal situation can be obtained. However, as we saw in §1.8, §1.9 and §2.2.3.2, certain numerical methods are motivated by the fact that low-rank approximations to high-frequency Helmholtz solutions do exist, for restricted domains or for high levels of absorption. The three situations we consider largely follow these patterns.

We use a parameter h in our description of the various situations. This is because we're interested in the implications of Theorem 3.2.3 for sweeping-type preconditioners. The preconditioners are constructed for particular discretisations of Helmholtz problems, on meshes of width h. As in §2.2.3.2, we are interested in applying our results to domains corresponding to blocks of Schur complement matrices (recall that Schur complements arise from decompositions of the discretisation matrices). Therefore we allow the dimensions of the domains (i.e. $b-a$ and d) to depend on h.

Note that in all of the lemmas, we take $h \sim k_R^{-\mu}$ where $1 \leq \mu \leq 2$. $\mu = 1$ corresponds to a grid with a fixed number of points per wavelength; $\mu = 3/2$ corresponds to a finer grid, which may be necessary as discussed in §1.4.2.

We simplify (3.1) in three different ways in three lemmas, as follows.

1) In Lemma 3.2.4 we make the left-hand side of (3.1) small, by considering the situation of a very narrow set of domains with $d \sim h$. Domains with $d \sim h$ are of interest; indeed recall that in §2.2.3.2 we recapped how Engquist and Ying used Theorem 2.2.22 with domains that had $d \sim h$ to create Theorem 2.2.23 and motivate a $\mathcal{H}$-matrix decomposition of the Schur complements. (In §4.2 we give our own versions of these results for Theorem 3.2.3.)

2) In Lemma 3.2.6 we make the right-hand side of (3.1) big, by considering a special form of absorption.

3) In Lemma 3.2.8 we again make the left-hand side of (3.1) small by considering $d \sim h$, but this time we also allow ε to decrease with increasing k_R, for a relatively small increase in the rank p.

This first lemma is for narrow domains with $d \sim h$ when ε is assumed to be independent of our parameters of interest (especially k).

Lemma 3.2.4. ***LHS small with narrow domains*** *Let $0 \leq k_I \leq k_R$ and $h \sim k_R^{-\mu}$ for $1 \leq \mu \leq 2$ and $d \sim h$ and $a \sim h^\nu$ for $0 \leq \nu \leq 1$ and $\nu < 2 - 1/\mu$. Then, given $\varepsilon \in (0,1)$, where ε' is independent of the other parameters of interest, there exists $k_0(\varepsilon) > 0$ such that (3.1) is satisfied for all $k_R \geq k_0(\varepsilon)$.*

The proof is contained in §3.4.6.

Remark 3.2.5. *Lemma 3.2.4 allows us to replace condition (3.1) in Theorem 3.2.3 by the conditions in Lemma 3.2.4 and the assumption that $k_R \geq k_0(\varepsilon)$ for some $\varepsilon \in (0,1)$.*

Note that, for whichever value of μ we choose in the range $(1,2]$, we are permitted to take $0 \leq \nu \leq 1$, corresponding to $h \lesssim a \lesssim 1$. The value $a = h$ is included in the possible range, a desirable small separation of only one grid width. Heuristically the small separation a is permitted because d is small: small-separation narrow domains are permissible, but fatter small-separation domains are not.

In the proof we use the fact that, under the conditions in Lemma 3.2.4 and Remark 3.2.5, the factor $k_R d^2/a \to 0$ algebraically in k_R, so that the left-hand side of (3.1)$\to 0$ as $k_R \to 0$ and so condition (3.1) is readily satisfied for $k_R \geq k_0(\varepsilon)$ for some $k_0(\varepsilon)$.

Lemma 3.2.4 and Remark 3.2.5 are valid for all $k_I \geq 0$. We now consider specifically the case when k_I grows with k_R. We choose $k_I = \beta k_R^\delta$, with some $\beta > 0$ and $0 < \delta \leq 1$, expecting that the Hankel function with this amount of damping of the oscillations admit separable expansions with less restrictive conditions on the domains in some way.

Lemma 3.2.6. ***RHS big with special form of absorption*** *Let the special form of absorption be $k_I = \beta k_R^\delta$ with some $\beta > 0$ and $0 < \delta \leq 1$. Let $h \sim k_R^{-\mu}$ with $1 \leq \mu \leq 2$ and $a \sim h^\nu$ with $0 \leq \nu \leq 1$. If $\tilde{\delta} := (\delta - \nu\mu)/2 > 0$, then*

given $\varepsilon \in (0,1)$*, where* $1/\varepsilon = \mathcal{O}(\exp(k_R^{\tilde{\delta}}))$*, there exists* $k_0(\varepsilon) > 0$ *such that* (3.1) *is satisfied for all* $k_R \geq k_0(\varepsilon)$*.*

The proof is contained in §3.4.6.

Remark 3.2.7. *Lemma 3.2.6 allows us to replace condition* (3.1) *in Theorem 3.2.3 by the conditions in Lemma 3.2.6 and the assumption that* $k_R \geq k_0(\varepsilon)$ *for some* $\varepsilon \in (0,1)$*.*

We see the factor $\exp(k_I a)$ on the right-hand side of (3.1) allows us to consider domains where d is larger, providing we have this special form of absorption where k_I grows with k_R. Allowing for larger d in this way comes at the expense of making the domains more separated: observe that effectively the conditions $a \sim h^\nu$ and $0 \leq \nu < \delta/\mu$ appear in Remark 3.2.7 instead of $\nu < 2 - 1/\mu$ in Remark 3.2.5. This δ/μ appears because we need $k_I a \to \infty$, as $k_R \to \infty$, for the $\exp(k_I a)$ factor on the right-hand side of (3.1) to grow exponentially as k_R increases. Then (3.1) is less restrictive and easier to satisfy as $k_R \to \infty$, allowing for larger values of d. (To see why the condition $0 \leq \nu < \delta/\mu$ makes $k_I a \to \infty$, as $k_R \to \infty$, recall that $k_I a = \beta k_R^\delta h^\nu = \beta k_R^\delta k_R^{-\mu\nu}$.)

The third lemma is for narrow domains with $d \sim h$ when ε is not k-independent, resulting in more separated domains than Lemma 3.2.4.

Lemma 3.2.8. ***LHS small and*** ε ***decreases with*** k_R***, with narrow domains*** *Let* $0 \leq k_I \leq k_R$ *and* $h \sim k_R^{-\mu}$ *with* $1 \leq \mu \leq 2$*,* $d \sim h$ *and* $a \sim h^\nu$ *with* $0 \leq \nu < 1 - 1/2\mu$*. Given* $\varepsilon' \in (0,1)$*, where* ε' *is independent of the other parameters of interest, then there exists* $k_0(\varepsilon')$ *such that for all* $k_R \geq k_0(\varepsilon')$*, there exist functions* $\{\phi_j, \chi_j\}_{j=1}^p$*, where*

$$p = \left\lceil C_2(\eta) \max\left\{1, \log^2\left(C_3(\eta)\max\left\{\frac{\exp(-2k_I a)k_R^{\mu\nu}}{\varepsilon'}, 1\right\}\right)\right\}\right\rceil, \tag{3.4}$$

and where C_2 *and* C_3 *depend only on* η*, such that*

$$\left| H_0(k\|x-y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y)\right| \leq \varepsilon' k_R^{-\mu\nu}, \tag{3.5}$$

for all $x \in D_1$*,* $y \in D_2$*. Observe that, compared to Theorem 3.2.3, in this result we take* $\varepsilon := \varepsilon' k_R^{-\mu\nu}$*.*

The proof is contained in 3.4.6.

The point of Lemma 3.2.8 is to consider the k-dependent value of ε, $\varepsilon := \varepsilon' k_R^{-\mu\nu}$. With the k-dependent value of ε the restrictive condition (3.1) in Theorem 3.2.3 becomes

$$\frac{k_R d^2}{a}\left[1+\left|\log\left(4(k_R a)^2+(k_R d)^2\right)\right|\right] \leq C_1 \varepsilon' k_R^{-\mu\nu}\exp(k_I a). \tag{3.6}$$

Then under the conditions in the lemma (3.6) is satisfied for $k_R \geq k_0(\varepsilon')$.

Note that the range of the k_R dependence for ε is $0 \leq \mu\nu \leq 3/2$.

We discuss the effect on the rank p later in §3.2.2.

We obtain Lemma 3.2.8 at the expense of making the domains more separated: compare the conditions $0 \leq \nu \leq 1-1/2\mu$ (from Lemma 3.2.8) and $0 \leq \nu < 2-1/\mu$ (from Lemma 3.2.4). This is due to the $k_R^{-\mu\nu}$ factor in ε, which results in the power of k_R being larger on the left-hand side of (3.1), so that the separation must increase to compensate; see the proof of Lemma 3.2.8 for more details.

3.2.2 Analysing the expression in (3.2) for the rank p

We firstly consider the expression in (3.2) for the rank p when ε is k-independent, i.e. in the cases of Remarks 3.2.5 and 3.2.7. To begin to understand (3.2) we observe that when $k_I = 0$ or constant relative to k_R, as $\varepsilon \to 0$ we see $p \sim \log^2(1/\varepsilon)$, as we expect due to this being the ε-dependence in the rank in Theorem 2.2.22. We next consider the case when ε is k-dependent in Lemma 3.2.8. Now when $k_I = 0$ or is constant relative to k_R, for p given by the expression in (3.4), we have that $p \sim \log^2(k_R^{\mu\nu}/\varepsilon')$ as $\varepsilon' \to 0$, i.e. we have acquired a $\log^2(k_R)$ dependence in the rank. However, for the $\log^2(k_R)$ increase in the rank, the quality of the approximation has increased algebraically, see (3.5).

In both (3.2) and (3.4), when ε is fixed and the term with the factor $\exp(-2k_I a)$ is dominating the inner maximum, we see that increasing $2k_I a$ is of benefit, as it causes the rank to decrease exponentially.

Note that here we only see benefits of increasing k_I when the term containing the factor $\exp(-2k_I a)$ dominates the inner maximum in (3.2) and (3.4), i.e. only for k_I sufficiently small. However, we expect benefits as we increase k_I, with no upper threshold on the value of k_I (within our range of values for k_I). In the next section we state a variant of the result where we see benefit in adding

absorption without an upper threshold on k_I (within our range of values for k_I). We note that improvements can only be seen if $k_I a$ increases as k_R increases, which for constant k_I and $a \sim h$, for example, is not the case. In §3.2.3.1 we discuss why a appears in the factor $\exp(-2k_I a)$ alongside k_I and also the cases where improvements are seen.

3.2.3 Improvements Due to Absorption

In §3.2.2 we see improvements due to absorption in the rank (3.2) of Theorem 3.2.3. We also see improvements due to absorption in the bigger right-hand side of (3.1) of Theorem 3.2.3, that allows the theorem to be valid for larger values of d, when the separation a is larger. However the tools in the proof of this result lend themselves to a second way of answering the question "as we increase absorption, what happens to the low-rank approximation?" Therefore we prove a slight variant on the original theorem which has statement as follows, showing improvement in the quality of the approximation rather than the rank.

Theorem 3.2.9. ***Variant of New Low-Rank Result*** *Let the domains and absorption be as defined in Definition 3.2.1 and Definition 3.2.2 respectively. Assume that for some constant $\eta > 0$, $\eta \text{dist}(D_1, D_2) > \text{diam}(D_1, D_2)$.*

Then there exists $C_1 > 0$ independent of a, b, d, k, and such that, given $\varepsilon \in (0, 1)$, if

$$\frac{k_R d^2}{a} \left[1 + \left|\log\left(4(k_R a)^2 + (k_R d)^2\right)\right|\right] \leq C_1 \varepsilon, \tag{3.7}$$

then there exist functions $\{\phi_j, \chi_j\}_{j=1}^{p}$ where

$$p = \left\lceil C_2(\eta) \max\left\{1, \log^2\left(\frac{C_3(\eta)}{\varepsilon}\right)\right\}\right\rceil, \tag{3.8}$$

and where C_2 and C_3 depend only on η, such that

$$\left|H_0(k\|x-y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y)\right| \leq \varepsilon \exp(-k_I a), \tag{3.9}$$

for all $x \in D_1, y \in D_2$.

The proof of this result is contained in §3.4.6.

The inequality (3.9) shows that the quality of the approximation improves exponentially as $k_I a$ increases in Theorem 3.2.9. Observe that in Theorem 3.2.9 the improvements due to absorption are in the quality of the approximation, whereas in Theorem 3.2.3 the potential improvements due to absorption are in the rank. In Theorem 3.2.9 there is also no upper threshold on the value of k_I for improvements due to absorption (within the range of values of k_I that we consider), whereas in Theorem 3.2.3 there is an upper threshold. In both theorems, improvements due to absorption are only seen for increasing k_R if $k_I a$ increases with k_R, we discuss this fact further in §3.2.3.1.

Note that similarly to (3.1) in Theorem 3.2.3, the condition (3.7) makes it difficult to see which domains Theorem 3.2.9 is valid for. So we use Lemma 3.2.4 and Lemma 3.2.8 to find two simpler sets of conditions that Theorem 3.2.9 is valid for, as follows.

Remark 3.2.10. *Note that the restrictive condition* (3.7) *in Theorem 3.2.9 is identical to the restrictive condition* (3.1) *in the original Theorem 3.2.3 when* $k_I = 0$. *Therefore Lemma 3.2.4 allows us to replace condition* (3.7) *in Theorem 3.2.9 by the conditions in Lemma 3.2.4 and the assumption that* $k_R \geq k_0(\varepsilon)$ *for some* $\varepsilon \in (0, 1)$.

Remark 3.2.11. *Note that the restrictive condition* (3.7) *in Theorem 3.2.9 is identical to the restrictive condition* (3.1) *in the original Theorem 3.2.3 when* $k_I = 0$. *Given* $\varepsilon' \in (0, 1)$, *where* ε' *is independent of the other parameters of interest, let* $\varepsilon := \varepsilon' k_R^{-\mu\nu}$. *Then the restrictive condition* (3.7) *in Theorem 3.2.9 becomes*

$$\frac{k_R d^2}{a}\left[1 + \left|\log\left(4(k_R a)^2 + (k_R d)^2\right)\right|\right] \leq C_1 \varepsilon' k_R^{-\mu\nu}. \tag{3.10}$$

Then, just as in the proof of Lemma 3.2.8, we can now show there exists $k_0(\varepsilon')$ *such that for all* $k_R \geq k_0(\varepsilon')$, (3.10) *is satisfied, and therefore allows us to replace condition* (3.7) *in Theorem 3.2.9 by the conditions in Lemma 3.2.8 and the assumption that* $k_R \geq k_0(\varepsilon')$. *Then, by substituting in the value of* ε, *the expression for the rank* p (3.8) *becomes*

$$p = \left\lceil C_2(\eta) \max\left\{1, \log^2\left(\frac{C_3(\eta) k_R^{\mu\nu}}{\varepsilon'}\right)\right\}\right\rceil, \tag{3.11}$$

and the separable expansion (3.9) *gives*

$$\left| H_0(k\|x-y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y) \right| \leq \varepsilon' \exp(-k_I a) k_R^{-\mu\nu}, \tag{3.12}$$

for all $x \in D_1$ *and* $y \in D_2$.

Note that we cannot apply Lemma 3.2.6 to Theorem 3.2.9 as it requires the $\exp(k_I a)$ factor on the right-hand side of (3.1), absent from (3.7).

Recall that the range of the k_R dependence for ε is $0 \leq \mu\nu \leq 3/2$.

3.2.3.1 Discussion of Results

Recall that the point of proving our new low-rank results is to find out what effect absorption has on the separable expansions of the Hankel functions (recall Rokhlin and Martinsson's Theorem 2.2.22 is only valid for real-valued wavenumbers). Stepping back and looking at the big picture, as k_I increases, the Hankel function $H_0(k\|x-y\|)$ has oscillations that get damped when absorption is included (recall the discussion in §1.9). Therefore, as k_I increases, we expect that approximating $H_0(k\|x-y\|)$ should get easier and that it should be possible to see the low-rank separable expansions getting 'easier' or approaching the ideal scenario in §3.2.1, for example by achieving a specified accuracy with separable expansions of increasingly low-rank. Indeed, in our results we see k_I appear beneficially in three places.

1) The first is in the decrease in the expression for the rank p (3.2) in Theorem 3.2.3, when the factor $\exp(-2k_I a)$ dominates the inner maximum.

2) The second is in making the right-hand side of the restrictive condition (3.1) bigger and so the condition (3.1) less restrictive. Consequently, in the case of Lemmas 3.2.4 and 3.2.8 and the resulting Remark 3.2.5, the sufficiently large $k_R \geq k_0(\varepsilon)$ needed to satisfy (3.1) and allow Theorem 3.2.3 to hold is smaller for $k_I > 0$. Also as a consequence of the right-hand side of (3.1) being bigger, in Lemma 3.2.6 and its resulting Remark 3.2.7, the height of the domain d can be much larger (compared with $d \sim h$ in Lemma 3.2.4 and 3.2.8) providing the separation a is larger.

3) The third is in the better quality of the approximation for the same rank with absorption seen in Theorem 3.2.9.

In all three places, k_I appears only in functions of $k_I a$. In fact, in all three places we have $\exp(-Ck_I a)$, with either $C = 1$ or $C = 2$. In order to benefit from this factor $\exp(-Ck_I a)$ as $k_R \to \infty$, we therefore need to find conditions that ensure $k_I a \to \infty$ as $k_R \to \infty$.

In fact, we already know a sufficient set of conditions from Lemma 3.2.6,

$$\begin{aligned} &\text{when } k_I = \beta k_R^{\delta} \text{ with some } \beta > 0,\ \delta > 0,\ h \sim k_R^{-\mu} \\ &\text{with } 1 \leq \mu \leq 2, \text{ and } a \sim h^{\nu} \text{ with } 0 \leq \nu \leq 1, \text{ if } \nu < \delta/\mu, \end{aligned} \tag{3.13}$$

then $k_I a \to \infty$ as $k_R \to \infty$.

Then providing (3.13) holds,

1) the rank p in (3.2) in Theorem 3.2.3 decreases exponentially as $k_R \to \infty$, up to the point where the term including $\exp(-2k_I a)$ no longer dominates the inner maximum.

2) the condition (3.1) in Theorem 3.2.3 gets exponentially easier to satisfy as $k_R \to \infty$, so that the sufficiently large k_R at which Lemmas 3.2.4 and 3.2.8 and the Remark 3.2.5 hold can be smaller. Also Lemma 3.2.6 and Remark 3.2.7 hold, with the additional condition on ε in the lemma statements and remarks; note that this condition on ε is easy to satisfy.

3) the quality of the approximation given by the right-hand side of (3.9) in Theorem 3.2.9 increases exponentially as $k_R \to \infty$.

Now, a natural question is, why are the benefits due to absorption seen only under the conditions of (3.13), that are a combination of the absorption k_I and the domains separation a? Note that the smaller a is, the closer the domain boxes D_1 and D_2 are to each other (see Definition 3.2.1) so the closer the argument of $H_0^{(1)}(k\|x-y\|)$ for $x \in D_1$ and $y \in D_2$ gets to the singularity at 0. The main benefit of absorption is that absorption damps the oscillations; absorption doesn't remove the singularity, see Figures 3-4, 3-5 and 3-6. Therefore, we may expect to only see improvements to the low-rank expansions when in the oscillatory regions away from the singularity, i.e. outside some separation $a > a_0$.

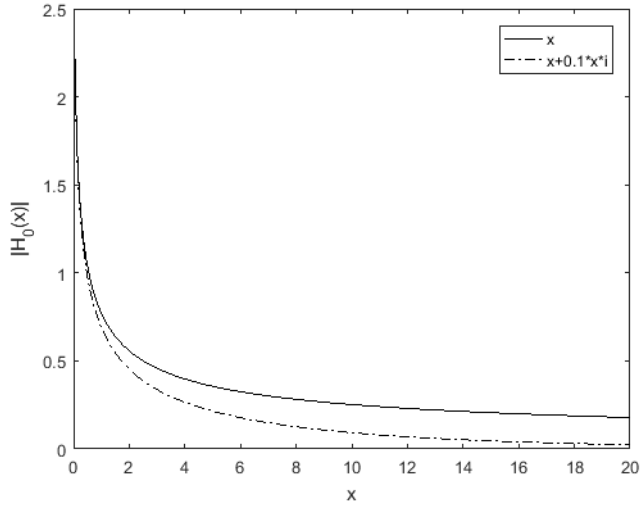


Figure 3-4: Plot of $|H_0(x)|$ against x. Note that $x \neq x \in D_1$, so $x = 1$ on this plot equates to real$(k\|x-y\|) = 1$ in our separable expansion results. Observe the singularity at $x = 0$ and the damping due to absorption.

In order to make this argument clearer, we examine the details of the damping rigorously. The exponential improvement for larger values of the argument is to be expected from analysis of the Hankel function. From [92, 10.2.5 with $v = 0$] we have:

$$H_0^{(1)}(z) \sim \sqrt{\frac{2}{\pi z}} \exp\left(\mathrm{i}z - \frac{\pi}{4}\right), \quad \text{as } z \to \infty, \text{ for } -\pi + \delta < \arg(z) \leq \pi. \quad (3.14)$$

We see that when z is real, the oscillations only decay with $\mathcal{O}(1/\sqrt{z})$ but for $z = z_R + \mathrm{i}z_I$ they decay with $\mathcal{O}(\exp(-z_I))$. This latter fact is also shown in the following lemma, which is easily obtained from other results gained during the proof of Theorem 3.2.3.

Lemma 3.2.12. *Let the domains be as in Definition 3.2.1. If $k_I = \beta k_R^\delta$ with some $\beta > 0$ and $0 < \delta \leq 1$, then*

$$\max_{\substack{x \in D_1 \\ y \in D_2}} |H_0(k\|x-y\|)| \leq C|\exp(-2\beta k_R^\delta a)| \left[\left|\log(4(k_R a)^2 + (k_R d)^2)^{\frac{1}{2}}\right| + 1\right]$$
$$\to 0, \text{ as } k_R^\delta a \to \infty.$$

The proof of this result is contained in §3.4.6.

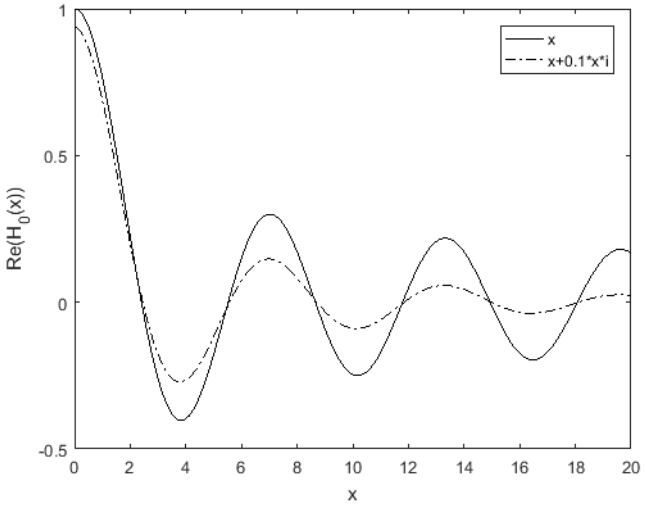


Figure 3-5: Plot of real($H_0(x)$) against x. The real part is oscillatory, observe the damping of the oscillations due to absorption.

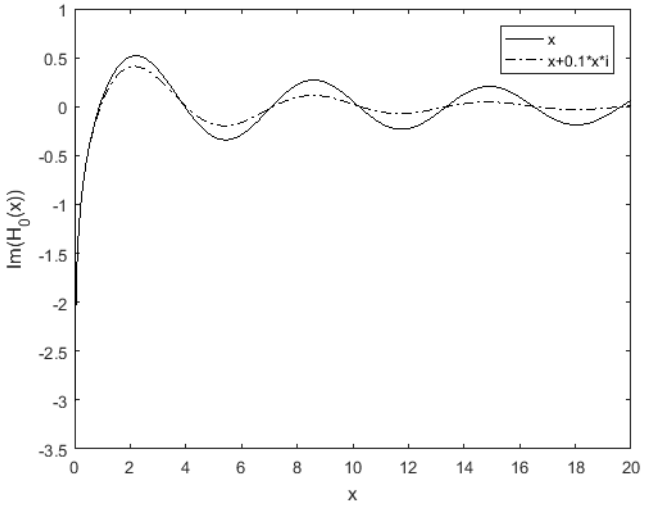


Figure 3-6: Plot of imag($H_0(x)$) against x. The imaginary part is oscillatory away from the singularity at $x = 0$. Observe how including absorption damps the oscillations, but makes little difference in the region $0 < x < 1$ near the singularity.

Since $\min\left(\mathrm{Im}(k\|x-y\|)\right) = 2ak_I$, the $\exp(k_I a)$ factor (seen repeatedly in points 1)-3) at the start of §3.2.3.1 above) has reappeared in Lemma 3.2.12, and Lemma 3.2.12 only shows improvements due to absorption when $k_I a = \beta k_R^{\delta} a \to \infty$.

Therefore, our new low-rank results (via points 1)-3) in §3.2.3.1), (3.14) and Lemma 3.2.12 all demonstrate that improvements due to absorption are seen due to the damping of the oscillations away from the singularity.

The need to be away from the singularity explains the need for $k_I a \to \infty$ as $k_R \to \infty$, as follows. The key point is that the effect of condition (3.13) (which is designed to make $k_I a \to \infty$ as $k_R \to \infty$) is that as $k_R \to \infty$ the minimum of the Hankel function's argument, i.e. $\min(k\|x-y\|)$ remains at some minimum distance from the singularity, or equivalently doesn't get closer to the singularity as $k_R \to \infty$. Thus (3.13) ensures that the separable expansion is applied in the oscillatory region away from the singularity and we see benefits due to absorption.

3.2.4 Comparison with Rokhlin and Martinsson's result

We compare Theorem 3.2.3 to Theorem 2.2.22 due to Rokhlin and Martinsson [82]. Their result only has real k, so we shall set our $k_I = 0$ in which case $k = k_R \in \mathbb{R}$.

We first compare the expressions for the rank p.

$$
\begin{aligned}
p \leq\ & \log(kb)|\log\varepsilon|^2 && \text{Theorem 2.2.22}\\
p =\ & \left\lceil C_2(\eta)\max\left\{1, \log^2\left(C_3(\eta)\max\left\{1, \frac{1}{\varepsilon}\right\}\right)\right\}\right\rceil && \text{Theorem 3.2.3}\\
p =\ & \left\lceil C_2(\eta)\max\left\{1, \log^2\left(C_3(\eta)\max\left\{1, \frac{k^{\mu\nu}}{\varepsilon'}\right\}\right)\right\}\right\rceil && \text{Lemma 3.2.8}\\
p =\ & \left\lceil C_2(\eta)\max\left\{1, \log^2\left(\frac{C_3(\eta)}{\varepsilon}\right)\right\}\right\rceil && \text{Theorem 3.2.9}\\
p =\ & \left\lceil C_2(\eta)\max\left\{1, \log^2\left(\frac{C_3(\eta)k^{\mu\nu}}{\varepsilon'}\right)\right\}\right\rceil && \text{Remark 3.2.11}
\end{aligned}
$$

Dependence on ε

Observe that in all the expressions for p, the dependence upon ε or ε' is identical as $\varepsilon \to 0$, or $\varepsilon' \to 0$, respectively, for fixed k.

Dependence on k

The expressions for p in our new Theorems 3.2.3 and 3.2.9 are independent of k, an improvement on Theorem 2.2.22 which allows logarithmic growth of k. The expressions in Lemma 3.2.8 and Remark 3.2.11 involve a $\log^2(k^{\mu\nu})$ factor, which may be better or worse than Theorem 2.2.22's $\log(kb)$ dependence, depending on the values of μ and ν. For example, $a \sim 1$ with $\nu = 0$ removes the dependence of $\log^2(k^{\mu\nu})$ on k, which is better than Theorem 2.2.22, whereas $a \sim h$ with $\nu = 1$ and $h \sim k^{-2}$ with $\mu = 2$ results in a $\log^2(k^2)$ dependence, which is worse than Theorem 2.2.22.

The expression for the rank in Theorem 2.2.22 has a log dependence on b, so that the longer the domains become, the higher the rank required. In contrast, our new Theorems 3.2.3 and 3.2.9 and their Lemma 3.2.8 and Remark 3.2.11 all have a dependence on the admissibility constant η, which is determined by the relationship between b, d and a, so that rank is not affected by an increase in b when a is increased proportionally, to maintain a constant value of η (recall that the admissibility condition is $\eta\text{dist}(D_1, D_2) > \text{diam}(D_1, D_2)$, or equivalently $2\eta a > \sqrt{(b-a)^2 + d^2}$).

Restrictions on the domains

We now compare the restrictions on the domains given in these conditions:

$ka > C(d)\lvert\log\varepsilon\rvert$	Theorem 2.2.22
$\left[1 + \left\lvert\log\left(4(ka)^2 + (kd)^2\right)\right\rvert\right] \dfrac{kd^2}{a} \le C_1\varepsilon$ $\eta\,\text{dist}(D_1, D_2) = 2\eta a$ $> \text{diam}(D_1, D_2) = \sqrt{(b-a)^2 + d^2}$	Theorem 3.2.3, Theorem 3.2.9.
$d \sim h, h \sim k^{-\mu}$ with $1 \le \mu \le 2$, $a \sim h^\nu$, with $0 \le \nu \le 1$ and $0 \le \nu < 2 - 1/\mu$	Remark 3.2.5
$d \sim h, h \sim k^{-\mu}$ with $1 \le \mu \le 2$, $a \sim h^\nu$, with $0 \le \nu < 1 - 1/2\mu$	Lemma 3.2.8

We observe that the conditions in Theorems 3.2.3 and Theorem 3.2.9 are explicit in d whereas the one in Theorem 2.2.22 isn't.

We also note that for larger k the condition in Theorem 2.2.22 becomes less restrictive, whereas the first condition in Theorems 3.2.3 and 3.2.9 becomes more restrictive. However, by Remark 3.2.5 and Lemma 3.2.8 we found alternative conditions that become less restrictive for larger k, making this aspect of our conditions comparable. The alternative conditions in Remark 3.2.5 and Lemma 3.2.8 allow for larger k_R by having the restriction $d \sim h$, but this restriction is comparable to Theorem 2.2.22 also. It is comparable because Theorem 2.2.22 has another implicit assumption that the domains are long and thin, as [82] is about scattering problems for elongated (i.e. long and thin) structures, and [82, Remark 2] says that the rank p increases rapidly with increasing box-height d.

Finally we compare the conditions on a and b. The ranges of the a and b values are comparable in the following sense: a and b can be $\mathcal{O}(h)$ size or many wavelengths (i.e. $\mathcal{O}(1)$) in all the theorems and remarks. However the condition $\eta \text{dist}(D_1, D_2) > \text{diam}(D_1, D_2)$ means that the combinations of a and b values we can have in Theorems 3.2.3 and 3.2.9 and Remark 3.2.5 and Lemma 3.2.8 are more restricted. Consider for example the case where $b = \mathcal{O}(1)$, $d = \mathcal{O}(h)$ and $a = \mathcal{O}(h)$, this case is permitted in Theorem 2.2.22, but $\eta\,\mathcal{O}(h) \not> \mathcal{O}(1)$ as $k \to \infty$, for any η constant with respect to k and so $\eta \text{dist}(D_1, D_2) > \text{diam}(D_1, D_2)$ is not satisfied. So the length and height b and d must be proportional to the separation a in our new results Theorems 3.2.3 and 3.2.9 and Remark 3.2.5 and Lemma 3.2.8 and they do not need to be proportional in Theorem 2.2.22. The need for proportionality in our result does mean that we cannot apply Theorems 3.2.3 and 3.2.9 and Remark 3.2.5 and Lemma 3.2.8 to the same domains with $a = \mathcal{O}(h)$ and $b = \mathcal{O}(1)$ as Engquist and Ying apply Theorem 2.2.22 to in Theorem 2.2.23, when they get results about the $\mathcal{H}$-matrix approximation of the Schur complement matrices. However, in the next chapter we see that we can still use our new low-rank results to prove something about the $\mathcal{H}$-matrix approximation of the Schur complement matrices, when they are approximated in a different version of the Hierarchical Matrix Framework to the one Engquist and Ying use in Theorem 2.2.23. We highlight that this different version is actually the version Engquist and Ying used in the numerical experiments in [34].

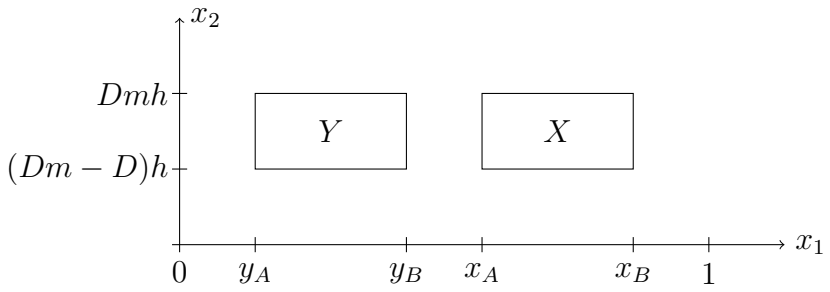


Figure 3-7: Domains for the Green's function

3.3 Statement and Analysis of New Low-Rank Results for the Green's Function

We now convert the low-rank results in §3.2 into low-rank results for the Green's functions of a series of half-plane problems. First we define the domains of these new half-plane problems, then we specify the half-plane problems and their Green's functions, and finally we present the new versions of the low-rank results for the Green's functions.

We find it advantageous to define the domains of the Green's functions to cover sections of Ω_m (see Definition 2.2.1) because in the next chapter these results are used to prove low-rank results about off-diagonal blocks of the $\mathbb{G}^m$ matrices formed by point-wise evaluations of the Green's functions on Ω_m (see Definition 2.2.7). Consequently, we define the domains in terms of the variables m, M, D and h, essential in determining the location and dimension of sections of Ω_m, see Chapter 2, Definition 2.2.2 and Assumption 2.2.3.

Definition 3.3.1. ***Notation for Green's functions domains and points*** *Given $m \in \{1, \ldots, M\}$, $D \geq 1$, $h > 0$ and $[x_A, x_B]$, $[y_A, y_B]$ non-overlapping intervals in $[0, 1]$, we define $d := Dh$ and*

$$X = [x_A, x_B] \times [Dmh - d, Dmh],$$

and

$$Y = [y_A, y_B] \times [Dmh - d, Dmh],$$

see for example Figure 3-7. Assuming, without loss of generality, that $[x_A, x_B]$ is

the right-most domain, we define

$$a := \frac{(x_A - y_B)}{2} \quad \textit{and} \quad b := \frac{(x_B - y_A)}{2},$$

which gives values indicating the separation and lengths of domains as in Definition 3.2.1 of the domains for the Hankel functions. Let $x = [x_1, x_2]^T \in X$ *and* $y = [y_1, y_2]^T \in Y$. *Let* $\text{diam}(X, Y) := \max\{\text{diam}(X), \text{diam}(Y)\}$ *where* $\text{diam}(X) = \text{diam}(Y) = \sqrt{(b-a)^2 + d^2}$ *and* $\text{dist(X, Y)} = 2a$.

Problem 3.3.2. ***Half-plane problems (including Green's functions and domains)*** *Let* f *have compact support. Let* u^m *be the solution of the Helmholtz equation*

$$\Delta u^m(x) + k^2 u^m(x) = -f(x), \quad x \in \Lambda_m, \tag{3.15}$$

where $k = k_R + \mathrm{i}k_I$, f *is the source* f *from the model with argument restricted to* Λ_m *and* $\Lambda_m := (-\infty, -\infty) \times (-\infty, (Dm+1)h]$, *satisfying*

$$u^m(x) = 0, \textit{ for all } x \in \partial\Lambda_m = (-\infty, \infty) \times [(Dm+1)h], \tag{3.16}$$

with the Sommerfeld radiation condition (SRC)

$$\frac{x}{\|x\|} \cdot \nabla u^m(x) - \mathrm{i}k u^m(x) = o\left(\frac{1}{\|x\|}\right) \textit{ as } \|x\| \to \infty. \tag{3.17}$$

The solution to these half-plane problems is the Green's function

$$G^m(x, y) := \frac{\mathrm{i}}{4} H_0(k\|x - y\|) - \frac{\mathrm{i}}{4} H_0(k\|x \ - M(y)\|),$$

where $m \in \{1, \ldots, M\}$ *and* $M(y)$ *reflects the point* y *in the line* $\partial\Lambda_m$ *(see Figure 2-7), and* H_0 *is the Hankel function of the first kind, see §1.1.1.1. The domains are as in Definition 3.3.1.*

The absorption is as defined in §1.9 and in §3.2. The sequence of half-plane problems are nearly identical to those in Definition 2.2.18, but with constant wavespeed and the source f being the source from Definition 1.1.2 on the whole of Λ_m, rather than spatially-varying wavespeed and source f^m only taking non-zero values on Ω_m. The Green's functions G^m are as in Definition 2.2.6 and Proposition 2.2.20 proves that the functions G^m are the Green's functions for the

half-plane problems in Definition 2.2.18 and therefore our half-plane problems in Problem 3.3.2.

Theorem 3.3.3. ***New Low-Rank Result for the Green's Function*** *We consider the half-plane problem (including domains and Green's function G^m) as defined in Problem 3.3.2. Then, providing $\eta \text{dist}(X,Y) > \text{diam}(X,Y)$ for some $\eta > 0$, there exists $C_4 > 0$ that is independent of m, h, a, b, k, η and d, such that, given $\varepsilon \in (0,1)$, if*

$$\frac{k_R(2(d+h))^2}{a}\left[1 + \left|\log(4(k_R a)^2 + (k_R 2(d+h))^2)\right|\right] \leq C_4 \varepsilon \exp(k_I a), \tag{3.18}$$

then there exist functions $\{\Phi_j, \Psi_j\}_{j=1}^R$ where

$$R = \left\lceil C_2(\eta) \max\left\{1,\, \log^2\left(C_5(\eta)\max\left\{\frac{\exp(-2k_I a)}{\varepsilon},\, 1\right\}\right)\right\}\right\rceil, \tag{3.19}$$

and where C_2 and C_5 depend only on η, such that

$$\left|G^m(x,y) - \sum_{j=1}^{R} \Phi_j(x)\Psi_j(y)\right| \leq \varepsilon, \tag{3.20}$$

for all $x \in X$ and $y \in Y$.

The proof is contained in §3.5

Remark 3.3.4. *When $d \sim h$, the restrictive condition* (3.18) *is equivalent to the restrictive condition* (3.1) *in Theorem 3.2.3, up to the constant on the right-hand side, which does not depend on the parameters of interest. Therefore Lemma 3.2.4 allows us to replace condition* (3.18) *in Theorem 3.3.3 by the conditions in Lemma 3.2.4 and the assumption that $k_R \geq k_0(\varepsilon)$ for some $\varepsilon \in (0,1)$. Note that the condition $d \sim h$ implies that D in Definition 3.3.1 satisfies $D = \mathcal{O}(1)$.*

Remark 3.3.5. *If we impose the condition $d \gtrsim h$, the restrictive condition* (3.18) *is equivalent to the restrictive condition* (3.1) *in Theorem 3.2.3, up to the constant on the right-hand side, which does not depend on the parameters of interest. Therefore Lemma 3.2.6 allows us to replace condition* (3.18) *in Theorem 3.3.3 by the conditions in Lemma 3.2.6 and the assumption that $k_R \geq k_0(\varepsilon)$ for some $\varepsilon \in (0,1)$. Note that the condition $d \sim h$ implies that D in Definition 3.3.1 satisfies $D \gtrsim 1$.*

Remark 3.3.6. *When $d \sim h$, the restrictive condition* (3.18) *is equivalent to the restrictive condition* (3.1) *in Theorem 3.2.3 up to the constant on the right-hand side, which does not depend on the parameters of interest. Given $\varepsilon' \in (0,1)$, where ε' is independent of the other parameters of interest, let $\varepsilon := \varepsilon' k_R^{-\mu\nu}$. Then the restrictive condition* (3.18) *in Theorem 3.3.3 becomes*

$$\frac{k_R d^2}{a}\left[1 + \left|\log(4(k_R a)^2 + (k_R d)^2)\right|\right] \leq C_6 \varepsilon' k_R^{-\mu\nu} \exp(k_I a). \tag{3.21}$$

Then, just as in the proof of Lemma 3.2.8, we can now show there exists $k_0(\varepsilon')$ such that for all $k_R \geq k_0(\varepsilon')$, (3.21) *is satisfied, and therefore allows us to replace condition* (3.18) *in Theorem 3.3.3 by the conditions in Lemma 3.2.8 and the assumption that $k_R \geq k_0(\varepsilon')$. Then, by substituting in the value of ε, the expression for the rank R* (3.19) *becomes*

$$R = \left\lceil C_2(\eta) \max\left\{1, \log^2\left(C_5(\eta)\max\left\{\frac{\exp(-2k_I a)k^{\mu\nu}}{\varepsilon'}, 1\right\}\right)\right\}\right\rceil, \tag{3.22}$$

and the expression (3.20) *for the separable expansion becomes*

$$\left|G^m(x,y) - \sum_{j=1}^{R} \Phi_j(x)\Psi_j(y)\right| \leq \varepsilon' k^{-\mu\nu}, \tag{3.23}$$

for all $x \in X$ and $y \in Y$. Note that the condition $d \sim h$ implies that D in Definition 3.3.1 satisfies $D = \mathcal{O}(1)$.

Recall that we had a second low-rank result (Theorem 3.2.9) showing the benefits of absorption in a different way i.e. that you get a better quality of approximation for the same rank with absorption than without. We now produce the analogue for the Green's function. We note that the conditions for Theorem 3.2.9 are identical to the conditions in the original Hankel function result Theorem 3.2.3, apart from the absence of the factor $\exp(k_I a)$ on the right-hand side of (3.7) compared to (3.1) and so it is possible to obtain a Green's function result using Theorem 3.2.9 using the method used to prove Theorem 3.3.3.

Theorem 3.3.7. ***Variant of New Low-Rank Result for the Green's Function*** *We consider the half-plane problem (including domains and Green's*

function G^m) defined by Problem 3.3.2. We further assume that $h \sim k_R^{-\mu}$ with $1 \leq \mu \leq 2$ and $a \sim h^\nu$ with $0 \leq \nu < 1$. Then providing $\eta \mathrm{dist}(X, Y) > \mathrm{diam}(X, Y)$ for some $\eta > 0$, there exists $C_4 > 0$ that is independent of a, b, k, η, d, m and h, such that, given $\varepsilon \in (0, 1)$, if

$$\frac{k_R(2(d+h))^2}{a}\left[1 + |\log(4(k_R a)^2 + (k_R 2(d+h))^2)|\right] \leq C_4\varepsilon, \tag{3.24}$$

then there exist functions $\{\Phi_j, \Psi_j\}_{j=1}^R$ where

$$R = \left\lceil C_2(\eta) \max\left\{1, \log^2\left(\frac{C_5(\eta)}{\varepsilon}\right)\right\}\right\rceil, \tag{3.25}$$

and where C_2 and C_5 depend only upon η, such that

$$\left|G^m(x, y) - \sum_{j=1}^{R} \Phi_j(x)\Psi_j(y)\right| \leq \varepsilon \exp(-k_I a), \tag{3.26}$$

for all $x \in X$ and $y \in Y$.

Remark 3.3.8. *As in §3.2, it is only possible to see improvements due to absorption in the limit $k_R \to \infty$ in the case with the special form of absorption $k_I = \beta k_R^\delta$, for some $\beta > 0$ and $0 < \delta \leq 1$, $\nu < \delta/\mu$ as this is the case for which $\varepsilon' \to 0$ as $k_R \to \infty$.*

This result follows from Theorem 3.2.9 in the same way that Theorem 3.3.3 followed from Theorem 3.2.3.

Remark 3.3.9. *When $d \sim h$, similarly to §3.2.3, we observe that the restrictive condition* (3.24) *is equivalent to the restrictive condition in the original Hankel function result* (3.1) *when $k_I = 0$, up to the constant on the right-hand side, which does not depend on the parameters of interest. Therefore Lemma 3.2.4 allows us to replace condition* (3.24) *in Theorem 3.3.7 by the conditions in Lemma 3.2.4 and the assumption that $k_R \geq k_0(\varepsilon)$ for some $\varepsilon \in (0, 1)$. Note that the condition $d \sim h$ implies that D in Definition 3.3.1 satisfies $D = \mathcal{O}(1)$.*

Remark 3.3.10. *When $d \sim h$, similarly to §3.2.3, we observe that the restrictive condition* (3.24) *is equivalent to the restrictive condition in the original Hankel function result* (3.1) *when $k_I = 0$, up to the constant on the right-hand side,*

which does not depend on the parameters of interest. Given $\varepsilon' \in (0,1)$*, where* ε' *is independent of the other parameters of interest, let* $\varepsilon := \varepsilon' k_R^{-\mu\nu}$*. Then the restrictive condition* (3.24) *in Theorem 3.3.7 becomes*

$$\frac{k_R d^2}{a}\left[1 + |\log(4(k_R a)^2 + (k_R d)^2)|\right] \leq C_6 \varepsilon' k_R^{-\mu\nu}. \tag{3.27}$$

Then, just as in the proof of Lemma 3.2.8, we can now show there exists $k_0(\varepsilon')$ *such that for all* $k_R \geq k_0(\varepsilon')$*,* (3.27) *is satisfied, and therefore allows us to replace condition* (3.24) *in Theorem 3.3.7 by the conditions in Lemma 3.2.8 and the assumption that* $k_R \geq k_0(\varepsilon')$*. Then, by substituting in the value of* ε*, the expression for the rank* R (3.25) *becomes*

$$R = \left\lceil C_2(\eta) \max\left\{1, \log^2\left(\frac{C_5(\eta) k_R^{\mu\nu}}{\varepsilon}\right)\right\}\right\rceil, \tag{3.28}$$

and the expression (3.26) *for the separable expansion becomes*

$$\left| G^m(x,y) - \sum_{j=1}^{R} \Phi_j(x)\Psi_j(y) \right| \leq \varepsilon \exp(-k_I a) k_R^{-\mu\nu}, \tag{3.29}$$

for all $x \in X$ *and* $y \in Y$*. Note that the condition* $d \sim h$ *implies that* D *in Definition 3.3.1 satisfies* $D = \mathcal{O}(1)$*.*

Note that, as in §3.2.3, we cannot apply Lemma 3.2.6 to (3.24) from Theorem 3.3.7, as it requires the $\exp(k_I a)$ factor on the right-hand side of (3.1), absent from (3.24).

3.4 Proof of New Low-Rank Result for the Hankel Function

3.4.1 Strategy of Proof of New Low-Rank Result

Before proving Theorem 3.2.3, i.e. the existence of a low-rank separable expansion for $H_0^{(1)}$, we must first prove some intermediate results. To explain the plan for

our intermediate results, we recall the following integral representation of $H_0^{(1)}(z)$:

$$H_0^{(1)}(z) = -\frac{2\mathrm{i}}{\pi}\exp(\mathrm{i}z)\int_0^\infty \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}}\,\mathrm{d}t, \quad \text{for } 0 < \mathrm{Re}(z) < \infty,\ \mathrm{Im}(z) = 0; \tag{3.30}$$

this can be found in [75, §4.1.2 (4.19)] due to [90, §7.13.3 (13.07)]. We then define

$$h_0^{(1)}(z) = -\frac{2\mathrm{i}}{\pi}\int_0^\infty \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}}\,\mathrm{d}t, \quad \text{for } 0 < \mathrm{Re}(z) < \infty, \tag{3.31}$$

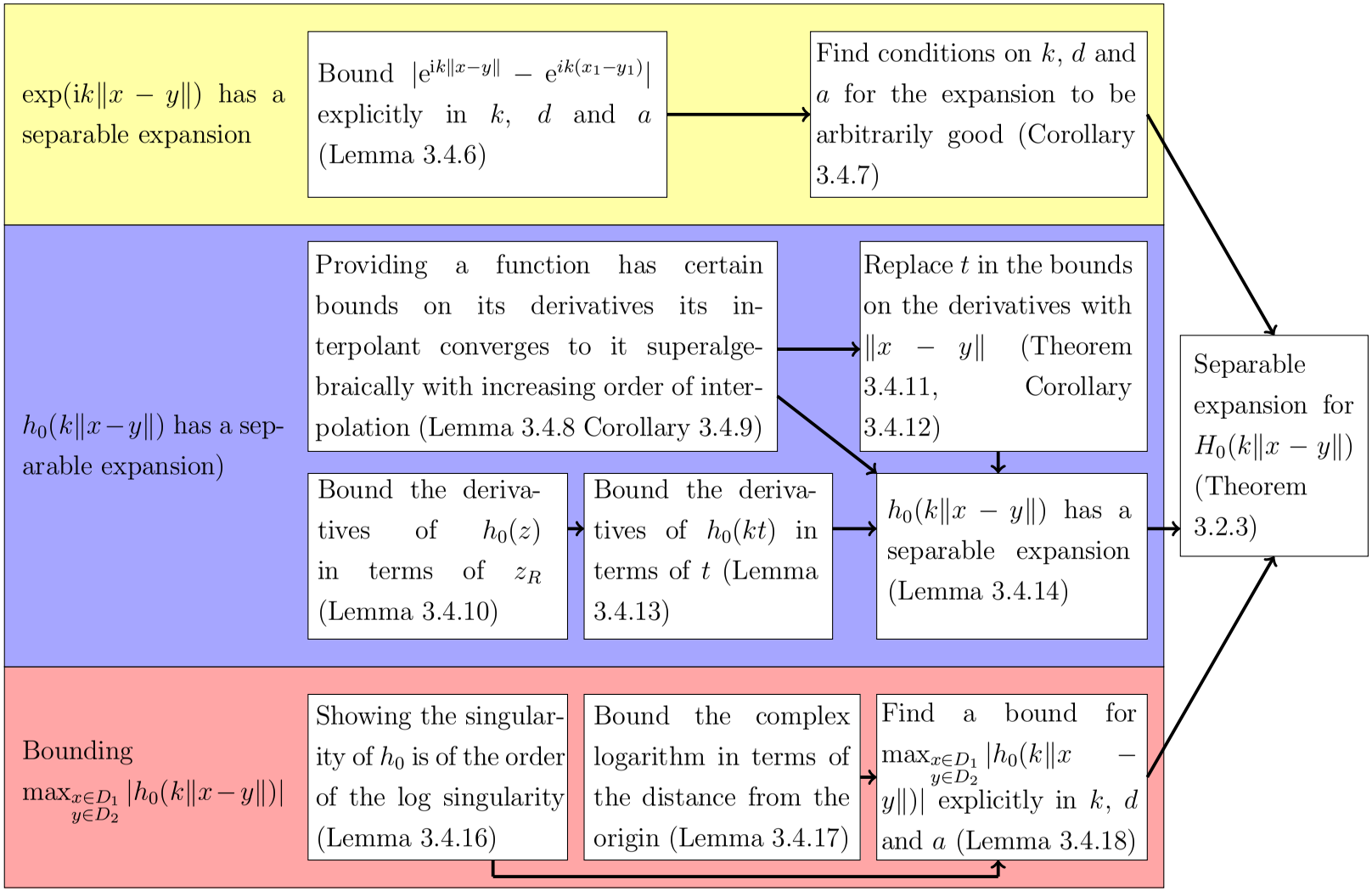
so that

$$H_0^{(1)}(z) = \exp(\mathrm{i}z)h_0^{(1)}(z). \tag{3.32}$$

We note that the expansion for $H_0^{(1)}(z)$ in (3.30) is only given in the classical literature for $z \in \mathbb{R}^+$ (where $\mathbb{R}^+$ denotes the positive real numbers and does not include 0), whereas to consider complex k we require the expansion to be valid in the quadrant with $0 < \mathrm{Re}(z) < \infty$ and $0 < Im(z) < \infty$. In §3.4.2 we prove that (3.30) is indeed valid in $0 < \mathrm{Re}(z) < \infty$, which contains the quadrant that we require.

The key step in the proof of Theorem 3.2.3 is that we use the theory of asymptotically smooth functions to get a k_R-independent separable expansion of $h_0^{(1)}(z)$ (in much of the rest of the thesis the superscript (1) is dropped for notational convenience). A short proof allows us to find a separable expansion for the part $\exp(\mathrm{i}z)$ as well, to get a separable expansion for $H_0^{(1)}(z)$ via (3.32).

So, in order to prove the main result, we let $z = k\|x-y\|$ in (3.32). As stated in the previous paragraph we find a separable expansion for h_0 using the theory of asymptotically smooth functions. We show that $\exp(\mathrm{i}k\|x-y\|)$ is close to the separable function $\exp(\mathrm{i}k(x_1-y_1))$ under certain conditions on k, d and a. Finally to combine these expansions and prove the existence of a low-rank separable expansion for H_0 via (3.32) we also need a bound on $\max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)|$. See the next page for a "roadmap" of the results.


$\exp(ik\|x-y\|)$ has a separable expansion
Bound $|e^{ik\|x-y\|} - e^{ik(x_1-y_1)}|$ explicitly in k, d and a (Lemma 3.4.6)
Find conditions on k, d and a for the expansion to be arbitrarily good (Corollary 3.4.7)
$h_0(k\|x-y\|)$ has a separable expansion)
Providing a function has certain bounds on its derivatives its interpolant converges to it superalgebraically with increasing order of interpolation (Lemma 3.4.8 Corollary 3.4.9)
Replace t in the bounds on the derivatives with $\|x-y\|$ (Theorem 3.4.11, Corollary 3.4.12)
Bound the derivatives of $h_0(z)$ in terms of z_R (Lemma 3.4.10)
Bound the derivatives of $h_0(kt)$ in terms of t (Lemma 3.4.13)
$h_0(k\|x-y\|)$ has a separable expansion (Lemma 3.4.14)
Bounding $\max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)|$
Showing the singularity of h_0 is of the order of the log singularity (Lemma 3.4.16)
Bound the complex logarithm in terms of the distance from the origin (Lemma 3.4.17)
Find a bound for $\max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)|$ explicitly in k, d and a (Lemma 3.4.18)
Separable expansion for $H_0(k\|x-y\|)$ (Theorem 3.2.3)

3.4.2 Validity of Integral Representation for the Hankel Function

We prove that the expansion for $H_0^{(1)}(z)$ in (3.30) still holds for the wider range of values $0 < \mathrm{Re}(z) < \infty$. In order to prove that the expansion still holds for this wider range of values, we need the following standard result on analytical continuation.

Definition 3.4.1. ***(Region)*** *(For example, see [97, Definition 3.14].) A region is a non-empty subset of $\mathbb{C}$ which is connected, that is, which cannot be expressed as $G_1 \cup G_2$ where G_1 and G_2 are non-empty, open, and disjoint.*

Theorem 3.4.2. *(For example, see [97, Theorem 5.16].) Suppose G is a region and that f and g are analytic[1] functions in G and $f(z) = g(z)$ for all $z \in S \subseteq G$, where S has a limit point in G. Then $f \equiv g$ within G.*

Theorem 3.4.2 is a classical result, appearing in various forms in many complex analysis works, see for example [109, Theorem 4.11 'Analytic Continuation' and Theorem 4.12 'Uniqueness of Analytic Continuation'].

Note that Theorem 3.4.2 implies a surprising and strong conclusion. Where two functions in a region G are equal on merely a countable set of points $S \subset G$, where S contains one accumulation point of a sequence of points in S, then the analytic functions are equal in the whole of G. This type of conclusion does not hold for functions that are simply real-differentiable in $\mathbb{R}$, so it shows that complex-differentiability or analyticity in $\mathbb{C}$ is a very strong concept. Ultimately Theorem 3.4.2 depends upon the fact that an analytic function in $\mathbb{C}$ can always be expressed in terms of its Taylor series or power series [109, p139] [2, p153].

We give an example of a consequence of Theorem 3.4.2 that we are interested in. Where two functions that are analytic on the region $0 < \mathrm{Re}(z) < \infty$, and one function is given by an integral representation of the other on the open subset that is the positive real line $\mathbb{R}^+$ (not including zero), then the functions are equal on the region $0 < \mathrm{Re}(z) < \infty$. (The positive real line $\mathbb{R}^+$ clearly contains an infinite number of sequences of points and their accumulation points.) This case is an example of the wider consequence of Theorem 3.4.2, that, in various

[1] The statement of the theorem in Priestley [97] says holomorphic functions in G, but recall that in this context holomorphism is equivalent to analyticity, see, for example [97, p69].

situations satisfying the conditions in Theorem 3.4.2, it is possible to extend the range of validity of a definition of a function or an integral representation of a function. The process of extending the range of validity of analytic functions or their representations in this way is known as analytical continuation, see for example [2, p152], [109, p139], [102, §10.23 p143, §10.2 p149, and §13.4 p211].

In order to prove that the expansion for $H_0^{(1)}(z)$ still holds for the wider range of values, we also need the following standard result on the analyticity of integral functions.

Definition 3.4.3. *(Regular contour) Let a contour be given by a differentiable map $M_c : \mathbb{R} \to \mathbb{C}$. A contour is regular if its derivative never vanishes.*

Theorem 3.4.4. *(For example, see [109, Theorem 2.84 and Theorem 2.85].) Let C be a contour going to infinity, any bounded part of which is regular. Then on any bounded part of C, let $f(z, w)$ be a continuous function of the complex variables z and w, where z ranges over a region D and w lies on the contour C. Also on any bounded part of C, let $f(z, w)$ be an analytic function z in D, for every value of w on C. Also suppose that*

$$\int_C f(z, w)dw,$$

is uniformly convergent. Then

$$F(z) := \int_C f(z, w)dw,$$

is an analytic function of z in D.

Now we prove that the expansion for $H_0^{(1)}(z)$ still holds for the wider range of values.

Lemma 3.4.5. *For $0 < \mathrm{Re}(z) < \infty$,*

$$H_0^{(1)}(z) = -\frac{2\mathrm{i}}{\pi} \exp(\mathrm{i}z) \int_0^\infty \frac{\exp(-zt)}{t^{1/2}(t - 2\mathrm{i})^{1/2}} dt, \tag{3.33}$$

holds.

Sketch Proof of Lemma 3.4.5. Let

$$g(z) := -\frac{2\mathrm{i}}{\pi}\exp(\mathrm{i}z)\int_0^\infty \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}}dt.$$

Since $H_0^{(1)}(z) = g(z)$ for $z \in \mathbb{R}^+$, by [90, 7.13.3 (13.07)], if both $H_0^{(1)}(z)$ and $g(z)$ are analytic in $0 < \mathrm{Re}(z) < \infty$, then $H_0^{(1)}(z) = g(z)$ in $0 < \mathrm{Re}(z) < \infty$ by Theorem 3.4.2, thus proving the lemma. Therefore it suffices to show that $H_0^{(1)}(z)$ and $g(z)$ are analytic in $0 < \mathrm{Re}(z) < \infty$.

$H_0^{(1)}(z)$ is an analytic function of z in $0 < \mathrm{Re}(z) < \infty$

By [89, 9.1.3 p358], $H_0^{(1)}(z) = J_0(z) + \mathrm{i}Y_0(z)$, where $J_0(z)$ is defined as a power series:

$$J_0(z) = \sum_{l=0}^{\infty} \frac{(-1)^l(\frac{1}{4}z^2)^l}{l!\Gamma(l+1)}$$

[93, (10.2.2)]. We note that this power series is analytic for all $z \in \mathbb{C}$, since its radius of convergence is ∞, by the ratio test, for example, [97, Example 2.10]. Similarly [93, (10.8.2)]

$$\begin{aligned} Y_0(z) =& \frac{2}{\pi}\left(\ln\left(\frac{1}{2}z\right)+\gamma\right)J_0(z) + \frac{2}{\pi}\left(\frac{\frac{1}{4}z^2}{(1!)^2} - \left(1+\frac{1}{2}\right)\frac{(\frac{1}{4}z^2)^2}{(2!)^2}\right. \\ &\left. + \left(1+\frac{1}{2}+\frac{1}{3}\right)\frac{(\frac{1}{4}z^2)^3}{(3!)^2} - \ldots\right) \\ =:& \frac{2}{\pi}\left(\ln\left(\frac{1}{2}z\right)+\gamma\right)J_0(z) + \mathbb{P}(z). \end{aligned} \tag{3.34}$$

Since the radius of convergence of the power series $\mathbb{P}(z)$ is also infinite by [97, Theorem 2.12] and $Y_0(z)$ is otherwise the sum of analytic functions it is therefore analytic in $0 < \mathrm{Re}(z) < \infty$.

$g(z)$ is an analytic function of z in $0 < \mathrm{Re}(z) < \infty$

It is clear that $\exp(\mathrm{i}z)$ is analytic. Therefore it remains to show that

$$h_0^{(1)}(z) = \int_0^\infty \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}}dt, \tag{3.35}$$

is analytic. Note that the analyticity of $h_0^{(1)}(z)$ follows from classical, standard results about complex functions defined using integrals, recall Theorem 3.4.4. We

seek to apply Theorem 3.4.4 with $w = t$,

$$f(z,t) = \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}},$$

the contour C being parameterised as $t \in (0,\infty) \to \mathbb{C}$, $t \to t$ and D being the right half-plane $0 < \mathrm{Re}(z) < \infty$.

To do this we note that

$$\frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}},$$

is an analytic function of $z \in D$ for all $t \in C$ and also a continuous function of $z \in D$ and of $t \in C$. That $\int_C f(z,t)dt$ converges uniformly at 0 (as required by Theorem 3.4.4) follows by observing that

$$\left| \int_0^{a'} \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}} dt \right| \leq \int_0^{a'} \frac{1}{t^{1/2}(t-2\mathrm{i})^{1/2}} dt < C_0 \int_0^{a'} \frac{1}{t^{1/2}} dt < \infty$$

for some constants a' and C_0. That $\int_C f(z,t)dt$ converges uniformly at ∞ (as required by Theorem 3.4.4) follows by observing that for $\mathrm{Re}(z) \geq x_0$, for some $x_0 > 0$,

$$\left| \int_0^{\infty} \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}} dt \right| \leq \int_0^{\infty} \frac{\exp(-x_0 t)}{t^{1/2}(t-2\mathrm{i})^{1/2}} dt$$

so that on any compact interval in $(0,\infty)$ the integral converges uniformly.

□

3.4.3 $\exp(\mathrm{i}k\|x-y\|)$ has a separable expansion

Firstly we use Taylor series expansions of $\|x-y\|$ to find a bound on

$$|\exp(\mathrm{i}k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1))|$$

that is explicit in terms of k, d and a .

Lemma 3.4.6. *Given $d > 0$ and $0 < a < b$, define $\acute{D}_1 := [a,b] \times [d]$ and $\acute{D}_2 := [-b,-a] \times [0]$. Let $x = [x_1, x_2]^T \in \acute{D}_1$ and $y = [y_1, y_2]^T \in \acute{D}_2$ (see Figure 3-8) and k be as is Definition 3.2.2, then*

$$|\exp(\mathrm{i}k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1))| \leq \exp(-k_I 2a)\left[\frac{k_R d^2}{4a} + \frac{k_I d^2}{4a} + \frac{k_R^2 d^2}{32a^2}\right].$$

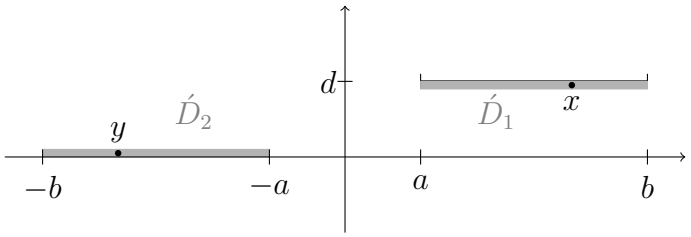


Figure 3-8: Line domains of Lemma 3.4.6.

Proof. Observe that

$$\|x-y\| = \sqrt{(x_1-y_1)^2+d^2} = (x_1-y_1)\sqrt{1+\frac{d^2}{(x_1-y_1)^2}}. \tag{3.36}$$

We seek to rearrange the expression for $\|x-y\|$ in (3.36) to find an expression for the difference between $\exp(\mathrm{i}k\|x-y\|)$ and its separable expansion $\exp(\mathrm{i}k(x_1-y_1))$; we then bound the difference to obtain the result. We apply Taylor's theorem (to first order) with $f(t)=(1+t)^n$, $n=1/2$ on $\left[0, \frac{d^2}{(x_1-y_1)^2}\right]$ (so $f'(t)=n(1+t)^{n-1}$) to find a bound on this factor. We get:

$$\sqrt{1+\frac{d^2}{(x_1-y_1)^2}} = 1+\frac{1}{2}\frac{1}{(1+\xi)^{1/2}}\frac{d^2}{(x_1-y_1)^2} \quad \text{for some } \xi \in \left(0, \frac{d^2}{(x_1-y_1)^2}\right).$$

This implies $\|x-y\| = (x_1-y_1)(1+\theta)$, where

$$\begin{aligned}\theta := \frac{1}{2}\left(\frac{1}{(1+\xi)^{1/2}}\frac{d^2}{(x_1-y_1)^2}\right) &\geq \frac{1}{2\left(\frac{(x_1-y_1)^2}{d^2}+\frac{(x_1-y_1)^4}{d^4}\right)^{\frac{1}{2}}} \\ &\geq \frac{1}{2\left(\frac{4b^2}{d^2}+\frac{16b^4}{d^4}\right)^{\frac{1}{2}}} \\ &\geq 0,\end{aligned} \tag{3.37}$$

where this inequality for θ and the following one (obtained by taking the maximum of the expression for θ in (3.37)) is useful later:

$$\theta \leq \frac{1}{2}\frac{d^2}{(x_1-y_1)^2}. \tag{3.38}$$

Now we use this expansion of $\|x-y\|$ in the exponential allows us to find an

equation containing our function and the separable expansion:

$$\begin{aligned}\exp(\mathrm{i}k\|x-y\|) &= \exp(\mathrm{i}k(x_1-y_1)(1+\theta)) \\ &= \exp(\mathrm{i}k(x_1-y_1))\exp(\mathrm{i}k(x_1-y_1)\theta) \\ &= \exp(\mathrm{i}k(x_1-y_1))\left(\exp(\mathrm{i}k(x_1-y_1)\theta)+1-1\right),\end{aligned}$$

rearranging this gives

$$\exp(\mathrm{i}k\|x-y\|)-\exp(\mathrm{i}k(x_1-y_1)) = \exp(\mathrm{i}k(x_1-y_1))\left(\exp(\mathrm{i}k(x_1-y_1)\theta)-1\right). \tag{3.39}$$

We denote the difference between $\exp(\mathrm{i}k\|x-y\|)$ and the separable expansion E and now seek an upper bound for E using (3.39):

$$\begin{aligned}E :&= \left|\exp(\mathrm{i}k\|x-y\|)-\exp(\mathrm{i}k(x_1-y_1))\right| \\ &= \left|\exp(\mathrm{i}k(x_1-y_1))\left(\exp(\mathrm{i}k(x_1-y_1)\theta)-1\right)\right|.\end{aligned}$$

Substituting in $k = k_R + k_I$, gives

$$E = \underbrace{\left|\exp(\mathrm{i}k_R(x_1-y_1))\right|}_{=1}\underbrace{\left|\exp(-k_I(x_1-y_1))\right|}_{\leq\exp(-k_I 2a)}\left|\exp(\mathrm{i}k(x_1-y_1)\theta)-1\right|. \tag{3.40}$$

Now substituting $k = k_R + k_I$ into the last factor in (3.40) and using $\exp(\mathrm{i}t) = \cos(t) + \mathrm{i}\sin(t)$ and the triangle inequality gives

$$\begin{aligned}E \leq \exp(-k_I 2a)\Big(&\Big|\underbrace{\exp(-k_I(x_1-y_1)\theta)\sin\left(k_R(x_1-y_1)\theta\right)}_{T_1}\Big| \\ &+\Big|\underbrace{\exp(-k_I(x_1-y_1)\theta)\cos\left(k_R(x_1-y_1)\theta\right)-1}_{T_2}\Big|\Big),\end{aligned}$$

by separating the real and imaginary parts of the last $|.|$. Now we seek an upper

bound on T_1 and T_2 in order to find an upper bound on E. We consider T_1 first:

$$\begin{aligned}
|T_1| &= \left| \exp(-k_I(x_1 - y_1)\theta) \sin\left(k_R(x_1 - y_1)\theta\right) \right| \\
&\leq \exp\left(-k_I a \times 0\right) \left| \sin\left(k_R(x_1 - y_1)\theta\right) \right| \quad \text{by (3.37)} \\
&\leq \left| \sin\left(k_R(x_1 - y_1)\theta\right) \right|,
\end{aligned}$$

We now apply Taylor's theorem (to first order) with $f(t) = \sin(t)$, on $[0, k_R(x_1 - y_1)\theta]$, to get

$$\sin\left(k_R(x_1 - y_1)\theta\right) = \cos(\xi')k_R(x_1 - y_1)\theta \quad \text{for some } \xi' \in (0, k_R(x_1 - y_1)\theta)\,.$$

Using the last equation and (3.38), we get the following bound for T_1 explicit in terms of k, d and a as follows

$$|T_1| \leq 1 \times 1 \times k_R(x_1 - y_1)\theta \leq k_R(x_1 - y_1)\frac{d^2}{2(x_1 - y_1)^2} = \frac{k_R d^2}{2(x_1 - y_1)} \leq \frac{k_R d^2}{4a}.$$

We now consider T_2. We apply Taylor's theorem (to second order) to $f(t) = \cos(t)$ on $[0, k_R(x_1 - y_1)\theta]$ to get

$$\cos(k_R(x_1 - y_1)\theta) = 1 - \frac{\cos(\xi'')}{2!}\left(k_R(x_1 - y_1)\theta\right)^2, \quad \text{for some } \xi'' \in (0, k_R(x_1 - y_1)\theta).$$

Substituting this bound in to the expression for T_2 gives

$$\begin{aligned}
|T_2| &= \left| \exp(-k_I(x_1 - y_1)\theta) \left(1 - \frac{\cos(\xi'')}{2!}\left(k_R(x_1 - y_1)\theta\right)^2\right) - 1 \right| \\
&\qquad \text{for some } \xi'' \in (0, k_R(x_1 - y_1)\theta), \\
&\leq |1 - \exp(-k_I(x_1 - y_1)\theta)| + \left| \frac{1}{2} k_R^2 (x_1 - y_1)^2 \theta^2 \right| |\exp(-k_I a \times 0)| \\
&\qquad \text{by (3.37)}, \\
&\leq |1 - \exp\left(-k_I(x_1 - y_1)\right)\theta| + \left| \frac{1}{2} k_R^2 (x_1 - y_1)^2 \theta^2 \right|.
\end{aligned}$$

In order to bound the first term on the right-hand side of this last inequality, we first note that

$$\max(\theta(x_1 - y_1)) \leq \frac{1}{2}\frac{d^2}{(x_1 - y_1)} \leq \frac{d^2}{4a}$$

by (3.38). Then, by Taylor's Theorem (to first order) with $f(t) = e^{-t}$ on $\left[0, k_I \frac{d^2}{4a}\right]$, we have

$$\exp\left(-k_I \frac{d^2}{4a}\right) = 1 - \exp(-\xi''')k_I \frac{d^2}{4a} \quad \text{for some } \xi''' \in \left(0, k_I \frac{d^2}{4a}\right).$$

Therefore

$$\begin{aligned} |T_2| \leq & \left|1 - \left(1 - \exp(\xi''')k_I \frac{d^2}{4a}\right)\right| + \left|\frac{1}{2}k_R^2 \left(\frac{d^2}{4a}\right)^2\right| \\ \leq & k_I \frac{d^2}{4a} + \frac{k_R^2 d^4}{32a^2}. \end{aligned}$$

Finally, combining our upper bounds for T_1 and T_2 we have

$$\begin{aligned} E \leq & \exp(-k_I 2a)\left(|T_1| + |T_2|\right) \\ = & \exp(-k_I 2a)\left[\frac{k_R d^2}{4a} + \frac{k_I d^2}{4a} + \frac{k_R^2 d^4}{32a^2}\right], \end{aligned}$$

which proves Lemma 3.4.6. □

The result Lemma 3.4.6 is not quite what is required, because the separable expansion needs to be arbitrarily close to the function. Hence in the next corollary we find a condition on k, d and a for this to be the case. The domains of x and y are also expanded to the full boxes required (i.e. from Figure 3-8 to Figure 3-3).

Corollary 3.4.7. *Let d, a, b, k, D_1, D_2, x and y be as in Definitions 3.2.1 and 3.2.2 and Figure 3-3. Given $\varepsilon \in (0,1)$, if $\frac{k_R d^2}{a} \leq e^{k_I a}\varepsilon$, then*

$$|\exp(\mathrm{i}k\|x - y\|) - \exp(\mathrm{i}k(x_1 - y_1))| \leq \varepsilon \quad \textit{for all } x \in D_1 \textit{ and } y \in D_2.$$

Proof. Recall $x = [x_1, x_2]$ and $y = [y_1, y_2]$. Since $x_2, y_2 \in [0, d]$, Lemma 3.4.6 applies for any value of $x \in D_1$ and $y \in D_2$ by the reflective symmetry of $\|.\|$.

Then, by Lemma 3.4.6 and the fact that $0 \leq k_I \leq k_R$, we have

$$
\begin{aligned}
&|\exp(\mathrm{i}k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1))| \\
&\leq \exp(-k_I 2a)\left[\frac{k_R d^2}{4a} + \frac{k_R d^2}{4a} + \frac{k_R^2 d^4}{32a^2}\right] \\
&\leq \exp(-k_I 2a)\exp(k_I a)\varepsilon\left[\frac{1}{4} + \frac{1}{4} + \frac{\exp(k_I a)\varepsilon}{32}\right], \quad \text{since } \frac{k_R d^2}{a} \leq \exp(k_I a)\varepsilon, \\
&\leq \varepsilon\left[\frac{1}{2}\exp(-k_I a) + \frac{1}{32}\varepsilon\right] \\
&\leq \varepsilon\left[\frac{1}{2} + \frac{1}{32}\right] \\
&\leq \varepsilon,
\end{aligned}
$$

as required. □

3.4.4 h_0 has a separable expansion

Next we use the theory of asymptotically smooth functions to find a separable expansion for h_0. For this we need the following result; we ultimately use this to show that the interpolant of $f(x) := h_0(k\|x-y\|)$ converges to f super-algebraically (with increasing order of interpolation). This interpolant is then demonstrated to be a separable function, which gives us the separable expansion of h_0.

Note the f used in this chapter is not the source term from the definition of the model problem Definition 1.1.2 in Chapter 1. Also in the next result i is an index, not the complex unit i and x is an arbitrary point in $\mathbb{R}^d$ (as opposed to a point in $\acute{D}_1$ from Figure 3-8 or D_1 from Definition 3.2.1).

Lemma 3.4.8. *Let $B = [a_1, b_1] \times \cdots \times [a_d, b_d]$ be a rectangle in d dimensions. Let $f \in C(B)$ be such that*

$$\left\| \frac{\partial^n}{\partial x_i^n} f \right\|_{\infty, B} \leq C_f n! \gamma_f^n, \tag{3.41}$$

for some C_f, for some γ_f, for all $n \in \mathbb{N}_0$, for all $i \in \{1, \ldots, d\}$ and $x \in \mathbb{R}^d$,

then

$$\|f - \mathcal{I}_B^p f\|_{\infty,B} \leq 4eC_f d(p+1)\left(2 + \frac{2}{\pi}\log(p+1)\right)^d \tag{3.42}$$

$$\times (1 + \gamma_f \text{diam}(B))\left(1 + \frac{2}{\gamma_f \text{diam}(B)}\right)^{-(p+1)}, \tag{3.43}$$

where

$$\mathcal{I}_B^p(x)f = \sum_{i_1,\ldots,i_d=0}^{p} f(x_{1,i_1},\ldots,x_{d,i_d})L_{1,i_1}(x_1)\ldots L_{d,i_d}(x_d),$$

where $L_{j,i}(x_j)$ $(0 \leq i \leq p)$ are one dimensional Lagrange polynomials and the interpolating nodes $x_{j,0},\ldots,x_{j,p}$ are the Chebyshev nodes defined as in Appendix B of [63].

To better understand Lemma 3.4.8, we first recall that functions are asymptotically smooth if their derivatives are only allowed to blow up in a controlled way – there is some variation in the definition of asymptotic smoothness in the literature, one example is the condition (3.41) that we use in Lemma 3.4.8. By looking at the form of the Taylor series expansion for a function and the form of the bounds on the derivatives for a function in the definition of asymptotic smoothness, we can see that asymptotic smoothness is "morally equivalent" to analyticity. (Indeed [63, Lemma E.5] gives a proof that a bi-variate asymptotically smooth function is analytic with respect to both its variables.) Once a function is proven to be a "nice" function in the sense that it is asymptotically smooth we can then use results from the theory of asymptotically smooth functions to find out further properties of these functions. Lemma 3.4.8 is one such result; it gives an error estimate for the approximation of asymptotically smooth functions with domains in d dimensions by their polynomial interpolant of degree p. We see that the last bracket dominates the rest of the right-hand side of (3.42) as $p \to \infty$. Note also that the rate of convergence of the right-hand side of (3.42) to 0 as $p \to \infty$ decreases as the diameter of the domain B increases. We expect this may happen because the function f is being evaluated over the bigger domain B and the approximation must also be done over the additional area. (However, this does not hold in all cases, for example in the case where $f(x) = x$, the extended domain makes no difference to the difficulty of the approximation by polynomial interpolants of degree > 1.)

We note that the constants C_f and γ_f depend upon the function f and therefore can depend upon our parameters of interest (recall from the statement of Theorem 3.2.3 that these include a, b, d, η and k). Therefore we keep track of the values of these constants in the rest of the section, where we prove the existence of the bounds on the derivatives, and hence find the values of the constants C_f and γ_f explicitly in the parameters of interest, see, for example, Lemmas 3.4.10, 3.4.13 and 3.4.14.

To prove Lemma 3.4.8 we adapt [63, Lemma B.7] (which is a similar result in d dimensions but the bound in (3.42) is less strict) to have the stricter bound of of [63, Theorem Satz 4.21] (which is a similar result in 1D). (This result and its proof are analogous to [6, Theorem 3.18], but this result uses a different definition of asymptotic smoothness which results in a different bound in (3.42).)

Proof. To begin with we follow Hackbusch's proof [63, Lemma B.7] to get (3.44)

$$\begin{aligned}
\|f - \mathcal{I}_B^p\|_{\infty,B} &= \left\| f - \prod_{j=1}^{d} \mathcal{I}_j^p f \right\|_{\infty,B} = \left\| \sum_{k=1}^{d} \left(\prod_{j=0}^{k-1} \mathcal{I}_j^p f - \prod_{j=1}^{k} \mathcal{I}_j^p f \right) \right\|_{\infty,B} \\
&= \left\| \sum_{k=1}^{d} \left(\prod_{j=1}^{k-1} \mathcal{I}_j^p \right) (f - \mathcal{I}_k^p f) \right\|_{\infty,B} \leq \sum_{k=1}^{d} \left\| \left(\prod_{j=1}^{k-1} \mathcal{I}_j^p \right) (f - \mathcal{I}_k^p f) \right\|_{\infty,B} \\
&\leq \sum_{k=1}^{d} \left(\prod_{j=1}^{k-1} C_{\text{stab}}^B(\mathcal{I}_j^p) \right) \|f - \mathcal{I}_k^p\|_{\infty,B} \,, \qquad (3.44)
\end{aligned}$$

where $(\mathcal{I}_k^p f)(x_1, \ldots, x_k, \ldots, x_d) = \sum_{i=0}^{p} f(x_1, \ldots, x_{k,i}, \ldots, x_d) L_{k,i}(x_k)$ for $k \geq 1$ and $(\mathcal{I}_0^p f) = f$ (i.e. $\mathcal{I}_0^p$ is the identity function).

We note for future reference that the Lebesgue constant $C_{\text{stab}}^B(\mathcal{I}_j^p) := \|\mathcal{I}_j^p\|_{C(B)\to C(B)}$ is equal to the Lebesgue constant $C_{\text{stab}}^{[a_j,b_j]}(\mathcal{I}_j^p) := \|\mathcal{I}_j^p\|_{C[a_j,b_j]\to C[a_j,b_j]}$ as the interpolant is only in the jth coordinate direction; both constants are just the operator norm over functions in the jth co-ordinate direction (on the interval $[a_j, b_j]$ in this case). So from now on we denote $C_{\text{stab}}(\mathcal{I}_j^p) := C_{\text{stab}}^B(\mathcal{I}_j^p) = C_{\text{stab}}^{[a_j,b_j]}(\mathcal{I}_j^p)$.

We now seek a bound on $\|f - \mathcal{I}_k^p\|_{\infty,B}$ (see (3.44)) to get the required result. By [63, Lemma B.3], using (3.41) we have the existence of a polynomial P_p of

degree p such that

$$\|f - P_p\|_{\infty,J} \leq 4\mathrm{e}C_f(1 + \gamma_f \operatorname{diam}(B))(p+1)\left(1 + \frac{2}{\gamma_f \operatorname{diam}(B)}\right)^{-(p+1)}, \quad (3.45)$$

where $J := [a_k, b_k]$ for any co-ordinate direction k and where we have used the fact that $\operatorname{diam}(J) \leq \operatorname{diam}(B)$. [63, Lemma B.3] is due to [10, Lemma 3.13], this uses the fact that (3.41) allows for holomorphic extension of the function, having this property then allows [27, §7.8 (8.7)] to be applied; this last result effectively says the function has a level of smoothness which allows the best approximating polynomial to converge superalgebraically to the function with polynomial order.

Now we recall that $\mathcal{I}_k^p$ interpolates polynomials of degree p exactly; then, by the definition of the Lebesgue constant,

$$\begin{aligned} \|f - \mathcal{I}_k^p\|_{\infty,J} &= \|(f - P_p) - \mathcal{I}_k(f - P_p)\|_{\infty,J} \\ &\leq (1 + C_{\mathrm{stab}}(\mathcal{I}_k^p))\|f - P_p\|_{\infty,J}, \end{aligned} \quad (3.46)$$

for all polynomials $P_p \in \mathcal{P}_p$, where $\mathcal{P}_p$ is the space of all polynomials of degree p. (This result comes from the definition of the Lebesgue constant and appears in [63, Lemma B.5] and [6, (3.37)]). Now recall that we are using Chebyshev nodes and so by [63, (B.13)]

$$C_{\mathrm{stab}}(\mathcal{I}_k^p) \leq 1 + \frac{2}{\pi}\log(p+1). \quad (3.47)$$

This last result appears in [63, (B.13)] (they say this is due to [98]) and [6, (3.40)] and Bebendorf in [6, p215] says additionally that this "is asymptotically optimal" in terms of p, which is due to the well-chosenness of Chebyshev nodes.

Thus combining (3.45), (3.46) and (3.47) we obtain

$$\begin{aligned} \|f - \mathcal{I}_B^p f\|_{\infty,J} \leq &\left(2 + \frac{2}{\pi}\log(p+1)\right) 4\mathrm{e}C_f(1 + \gamma_f \operatorname{diam}(B)) \\ &\times (p+1)\left(1 + \frac{2}{\gamma_f \operatorname{diam}(B)}\right)^{-(p+1)}. \end{aligned} \quad (3.48)$$

Recall that f is a function of $x \in \mathbb{R}^d$. Since the right-hand side of (3.48) is independent of x and k we can then take the supremum of each side over the

remaining co-ordinate directions without changing the bound, so (3.48) holds with J replaced by B. The required estimate then follows by combining (3.48) with (3.44) and applying (3.47) to $C^B_{\text{stab}}(\mathcal{I}^p_j) = C_{\text{stab}}(\mathcal{I}^p_j)$ in (3.44). □

Corollary 3.4.9. *Theorem 3.4.8 also holds for complex valued functions f with a constant of 8 instead of 4 in the right-hand side of* (3.42).

Proof. We assume we have (3.41) for our complex valued function f. Then we split f into its complex and imaginary parts like this: $f = f_1 + \mathrm{i}f_2$. We then note that

$$\left\| \frac{\partial^n}{\partial x_i^n} f_j \right\|_{\infty,B} \leq \left\| \frac{\partial^n}{\partial x_i^n} f \right\|_{\infty,B}, \text{ for } i \in \{1,\ldots,d\} \text{ and } j \in \{1,2\},$$

so that (3.41) also applies to both f_1 and f_2 individually; hence we can apply Theorem 3.4.8 to f_1 and f_2 to get the bound in (3.42) for f_1 and f_2. Then by the triangle inequality

$$\begin{aligned} \|f - \mathcal{I}^q_B f\|_{\infty,B} &= \|f_1 + \mathrm{i}f_2 - \mathcal{I}^q_B f_1 - \mathrm{i}\mathcal{I}^q_B f_2\|_{\infty,B} \\ &\leq \|f_1 - \mathcal{I}^q_B f_1\| + \|f_2 - \mathcal{I}^q_B f_2\|_{\infty,B}. \end{aligned}$$

Finally by applying the bound (3.42) for f_1 and f_2 to the right-hand side of this equation we obtain a bound for f of the form (3.42) but with the right-hand side multiplied by 2. □

To use this result to obtain a separable expansion of $h_0(k\|x-y\|)$, we first need to prove there are bounds of the form (3.41) for h_0. Note that, since k has a non-zero imaginary part, we need to consider the function $h_0(z)$ with z complex.

Lemma 3.4.10. *If $z = z_R + \mathrm{i}z_I$ where $z_R > 0$ and $z_R \geq z_I \geq 0$, then there exists $C > 0$ such that*

$$n = 0, \qquad \left| \left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^n h_0(z) \right| \leq C z_R^{-n} = C, \qquad z_R \in [1,\infty), \tag{3.49}$$

$$\text{for all } n \geq 1, \qquad \left| \left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^n h_0(z) \right| \leq C(n-1)! z_R^{-n}, \quad z_R \in (0,\infty). \tag{3.50}$$

Proof. Note that this proof uses a lot of ideas found in [75, Lemma 4.5], but the result is adapted to give bounds on the derivatives that are fully explicit in n and allow the argument z to be complex. We make use of the variant on the integral expansion seen earlier in (3.31), which we again note is found in [75, §4.1.2 (4.19)] due to [90, §17.13.3 (13.07)],

$$h_0(z) = \frac{-2\mathrm{i}}{\pi} \int_0^\infty \frac{\exp(-zt)}{t^{1/2}(t-2\mathrm{i})^{1/2}}\,\mathrm{d}t, \quad \text{for } 0 < \mathrm{Re}(z) < \infty,$$

reproduced here for clarity.

Now the complex derivative of $\exp(-tz)$ ($z \in \mathbb{C}$, $t \in \mathbb{R}$), is $-t\exp(-tz)$, so

$$\left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^n h_0(z) = (-1)^{n+1}\frac{2\mathrm{i}}{\pi}\int_0^\infty \frac{\exp(-zt)t^{n-1/2}}{(t-2\mathrm{i})^{1/2}}\,\mathrm{d}t, \qquad \text{for all } n \in \mathbb{N}_0.$$

Now

$$\left|\left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^n h_0(z)\right| = \left|\frac{2}{\pi}\int_0^\infty \frac{\exp(-z_R t)\exp(-z_I t\mathrm{i})t^{n-1/2}}{(t-2\mathrm{i})^{1/2}}\,\mathrm{d}t\right|, \qquad \text{for all } n \in \mathbb{N}_0,$$

by the inequality $\left|\int_a^b f(z)\,\mathrm{d}z\right| \le \int_a^b |f(z)|\,\mathrm{d}z$,

$$\left|\left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^n h_0(z)\right| \le \frac{2}{\pi}\int_0^\infty \left|\frac{\exp(-z_R t)t^{n-1/2}}{(t-2\mathrm{i})^{1/2}}\right|\,\mathrm{d}t, \qquad \text{for all } n \in \mathbb{N}_0.$$

Now by the triangle inequality we split this into two integrals over two parts of the domain

$$\left|\left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^n h_0(z)\right| \le \frac{2}{\pi}\big(|I_1(z)| + |I_2(z)|\big), \tag{3.51}$$

where

$$I_1(z) := \int_0^1 \left|\frac{\exp(-z_R t)t^{n-1/2}}{(t-2\mathrm{i})^{1/2}}\right|\,\mathrm{d}t$$

and

$$I_2(z) := \int_1^\infty \left|\frac{\exp(-z_R t)t^{n-1/2}}{(t-2\mathrm{i})^{1/2}}\right|\,\mathrm{d}t.$$

This is in order to make the change of variables $y = z_R t$ and make use of the fact that

$$|t-2\mathrm{i}|^{1/2} \ge \max\{t^{1/2}, 2^{1/2}\},$$

to obtain

$$\begin{aligned} |I_1(z)| &\leq \frac{1}{\sqrt{2}\, z_R^{n+1/2}} \int_0^{z_R} \exp(-y) y^{n-1/2} \,\mathrm{d}y, && \text{for all } n \in \mathbb{N}_0, \\ |I_2(z)| &\leq \frac{1}{z_R^n} \int_{z_R}^{\infty} \exp(-y) y^{n-1} \,\mathrm{d}y, && \text{for all } n \in \mathbb{N}_0, \end{aligned} \tag{3.52}$$

where the absolute value signs can be removed as the integrand is now real valued and positive.

Case $n \in \mathbb{N}$

Firstly we bound I_1. Note for all $y \in (0, z_R)$, $y^{-1/2} z_R^{-1/2} \leq y^{-1}$, so

$$|I_1(z)| \leq \frac{1}{\sqrt{2}\, z_R^n} \int_0^{z_R} \exp(-y) y^{n-1} \,\mathrm{d}y \leq \frac{1}{\sqrt{2}\, z_R^n} \int_0^{\infty} \exp(-y) y^{n-1} \,\mathrm{d}y.$$

Recall from the definition of the gamma function

$$\Gamma(n) = (n-1)! = \int_0^{\infty} y^{n-1} \mathrm{e}^{-y} \,\mathrm{d}y, \quad \text{for all } n \in \mathbb{N}, \tag{3.53}$$

and therefore

$$|I_1(z)| \leq \frac{1}{\sqrt{2} z_R^n} (n-1)!, \quad \text{for all } n \in \mathbb{N}. \tag{3.54}$$

Secondly we bound I_2.

$$|I_2(z)| \leq \frac{1}{z_R^n} \int_{z_R}^{\infty} \exp(-y) y^{n-1} \,\mathrm{d}y \leq \frac{1}{z_R^n} \int_0^{\infty} \exp(-y) y^{n-1} \,\mathrm{d}y,$$

so

$$|I_2(z)| \leq \frac{1}{z_R^n} (n-1)!, \quad \text{for all } n \in \mathbb{N}, \tag{3.55}$$

by (3.53). Combining (3.54) and (3.55) and (3.52) we obtain

$$\left| \left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^n h_0(z) \right| \leq C \frac{1}{z_R^n} (n-1)!, \quad \text{where} \quad C = \frac{2}{\pi} \left(\frac{1}{\sqrt{2}} + 1 \right), \quad \text{for all } n \in \mathbb{N}.$$

Case $n = 0$

For $z_R \in [1, \infty)$, clearly by (3.52) $I_1(z) \leq C_1 z_R^{-1/2}$, where $C_1 > 0$ and

$$I_2(z) \leq \int_1^{\infty} \frac{\exp(-y)}{y} \,\mathrm{d}y \leq C_2,$$

so overall the result holds with $C := \max\{C, C_1, C_2\}$. □

The next result, which we quote from [63], shows how asymptotic smoothness of $f(t)$ implies that $f(\|x-y\|)$ satisfies an analogous condition, which later helps us in converting our bounds on the derivatives of $h_0(z)$ into bounds on the derivatives of $h_0(k\|x-y\|)$.

Theorem 3.4.11. *[63, Theorem E.8 with $t \in (-d_f, 0)$ condition dropped, as explained in proof] If the function f is asymptotically smooth in the sense of [63, E.10a-c], i.e.*

$$\left|\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n f(t)\right| \leq Cn!n^p\gamma^n|t|^{-n-s}, \text{ for } t \in (0, d_f)\setminus\{0\},\ d_f > 0,\ n \in \mathbb{N},\ s \in \mathbb{R}, \tag{3.56}$$

then for all $\hat{\gamma} > 1$ there exists a $C_{\hat{\gamma}}$, such that for all directional derivatives D, the function $F(x,y) := f(\|x-y\|)$ satisfies

$$|D^n F(x,y)| \leq\ C_{\hat{\gamma}} n!\hat{\gamma}^n \|x-y\|^{-n-s}, \quad \text{where } 0 \neq \|x-y\| < d_f.$$

Sketch Proof. This is a sketch of how to alter the proof of [63, Theorem E.8] so that it doesn't require the condition $t \in (-d_f, 0)$. This sketch proof can only be understood in conjunction with the proof of [63, Theorem E.8] in [63]. (The changes are small, so it is not worthwhile to reproduce the proof in [63] here.) We proposed the change to drop the condition $(-d_f, 0)$ to Hackbusch and had it confirmed in a private communication with Hackbusch dated 24/06/2017. To explain the change, since $\|x-y\| > 0$ for all x and y, Theorem 3.4.11 only needs to be proved for $t = \|x-y\| \in (0, d_f)$. Part of the proof in [63, Theorem E.8] is to extend f analytically into the complex plane. Originally the extension was made on the unnecessarily large domain $t \in (-d_f, d_f)\setminus\{0\}$ and the result proved using f on that extended area around $t \in (-d_f, d_f)\setminus\{0\}$. But since $\|x-y\| > 0$, it is clear the extension can simply be made on the domain $t \in (0, d_f)$ instead and the result proved using f on that extended area around $t \in (0, d_f)$. □

Note that Theorem 3.4.11 holds for all $\hat{\gamma} > 1$ and that the constant $C_{\hat{\gamma}}$ depends only depends upon $\hat{\gamma}$ and C from (3.56).

We want to use Theorem 3.4.11 with $f(t) := h_0(kt)$, however since this is a complex valued function we must first prove the following Corollary.

Corollary 3.4.12. *Theorem 3.4.11 also holds for complex valued functions f.*

Proof. Split f into its complex and imaginary parts as follows $f = f_1 + if_2$, then note that $\left|\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n f_j(t)\right| \leq \left|\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n f(t)\right|$, for $n \in \mathbb{N}$ and $j \in \{1, 2\}$, so that (3.56) applies to f_1 and f_2. Hence we can apply Theorem 3.4.11 to f_1 and f_2 to get $F_j(x, y) := f_j(\|x - y\|)$ which satisfy

$$|D^n F_j(x, y)| \leq C_{\hat{\gamma}} n! \hat{\gamma}^n \|x - y\|^{-n-s}, \quad \text{where } 0 \neq \|x - y\| < d_f.$$

Then by the triangle inequality this last condition can also be obtained for $F(x, y) := f(\|x - y\|)$. □

Now all we need to do to apply Corollary 3.4.12 to $f(t) := h_0(kt)$ is to adapt the bounds on h_0's derivatives we have in (3.49) and (3.50) to prove (3.56) holds for our f. We note that the right-hand sides of the bounds (3.49) and (3.50) are already very similar to (3.56) with $p = 0$, $s = 0$, $\gamma = 1$, and $z_R = t$, but in the next lemma we get something that is an exactly comparable form to (3.56).

Lemma 3.4.13. *Let k be as in Definition 3.2.2 and let $f(t) := h_0(kt)$, then there exists $C > 0$ such that*

$$\left|\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n f(t)\right| \leq Cn!\sqrt{2}^n t^{-n}, \quad \textit{for } t \in \mathbb{R}^+,\ n \in \mathbb{N}. \tag{3.57}$$

Proof. By the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}t} h_0(kt) = k h_0'(z)\Big|_{z=kt}.$$

So

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n h_0(kt) = k^n \left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^n h_0(z)\Big|_{z=kt}.$$

Note that $|k^n| = |k|^n = (k_R^2 + k_I^2)^{n/2} \leq (2k_R^2)^{n/2} = \sqrt{2}^n k_R^n$ as $k_I \leq k_R$ so

$$\begin{aligned}
\left|\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n f(t)\right| &\leq \sqrt{2}^n k_R^n \left|\left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^n h_0(z)\Big|_{z=kt}\right| \\
&\leq \sqrt{2}^n k_R^n C(n-1)!(k_R t)^{-n}, \quad \text{by Lemma 3.4.10,} \\
&= C\sqrt{2}^n (n-1)! t^{-n},
\end{aligned}$$

so the result holds with $C = C$. $\square$

Finally we apply Corollary 3.4.12 to $f(t) := h_0(kt)$, to obtain bounds of the form (3.41) for our f, that allow us to apply Corollary 3.4.9, that ultimately yields a separable expansion for $h_0(k\|x-y\|)$.

Lemma 3.4.14. *Let the domains and absorption be as in Definitions 3.2.1 and 3.2.2 and Figure 3-3. Additionally we assume that the boxes are admissible in the following sense: for some constant $\eta > 0$, $\eta \mathrm{dist}(D_1, D_2) > \mathrm{diam}(D_1, D_2)$ (i.e. that they are well separated for their size). Then there exist $p_0(\eta)$, $C(\eta)$, $\widehat{C}(\eta)$ such that given $\varepsilon > 0$ there exist functions $\{\varphi_j, \chi_j\}_{j=0}^{p}$ such that*

$$\left| h_0(k\|x-y\|) - \sum_{j=1}^{p} \varphi_j(x)\chi_j(y) \right| \leq \varepsilon, \tag{3.58}$$

for all $x \in D_1$ and $y \in D_2$ and the number of functions p satisfies

$$p \geq \max\left\{ p_0(\eta), \frac{1}{\widehat{C}(\eta)} \log^2\left(\frac{C(\eta)}{\varepsilon}\right) \right\}.$$

Thus, in particular, $p = \mathcal{O}\left(\log^2\left(\frac{1}{\varepsilon}\right)\right)$ as $\varepsilon \to 0$.

Proof. Let $f(t) := h_0(kt)$, then by Lemma 3.4.13,

$$\left| \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^n f(t) \right| \leq \; Cn!\sqrt{2}^n t^{-n}, \qquad t \in \mathbb{R}^+, \; n \in \mathbb{N}. \tag{3.59}$$

Note that (3.59) is the condition (3.56) required for Corollary 3.4.12 with $\gamma = \sqrt{2}$, $p = 0$, $s = 0$ and we can take the maximum required value of t as d_f, i.e. $d_f := \sqrt{4b^2 + d^2}$. Applying Corollary 3.4.12 we then have for all $\hat{\gamma} > 1$,

$$\left| \frac{\partial^n}{\partial x_j^n} h_0(k\|x-y\|) \right| \leq \; C_{\hat{\gamma}} n! \hat{\gamma}^n \|x-y\|^{-n}, \quad \text{where } 0 \neq \|x-y\| < d_f, \text{ for } j = \{1,2\},$$

where $C_{\hat{\gamma}}$ does not depend on any parameters except C from (3.59) and $\hat{\gamma}$. Now define $B \colon = [a,b] \times [0,d] = D_1$. We wish to apply Corollary 3.4.9 to $f(x) :=$

$h_0(k\|x-y\|)$ on B. So we take the ∞-norm over B for our last bound to get

$$\left\|\frac{\partial^n}{\partial x_j^n}f(x)\right\|_{\infty,B} \leq \max_{\substack{x\in D_1\\ y\in D_2}}\left(C_{\hat{\gamma}}n!\left(\frac{\hat{\gamma}}{\|x-y\|}\right)^n\right) = C_{\hat{\gamma}}n!\left(\frac{\hat{\gamma}}{2a}\right)^n, \quad \text{for } j=\{1,2\},$$

so the bounds needed for Corollary 3.4.9 hold with $C_f := C_{\hat{\gamma}}$ and $\gamma_f := \frac{\hat{\gamma}}{2a}$. Thus applying Corollary 3.4.9 we get the following bound:

$$\begin{aligned}\|f-\mathcal{I}_B^q f\|_{\infty,B} \leq 8eC_{\hat{\gamma}}2 &\left(1+\frac{2}{\pi}\log(q+1)\right)^2(q+1)\left(2+\frac{2}{\pi}\log(q+1)\right)\\ &\times\left(1+\frac{\hat{\gamma}\,\mathrm{diam}(B)}{2a}\right)\left(1+\frac{2\times 2a}{\hat{\gamma}\,\mathrm{diam}(B)}\right)^{-(q+1)},\end{aligned} \tag{3.60}$$

where $\mathrm{diam}(B)=\sqrt{(b-a)^2+d^2}$, see Definition 3.2.1. Note that (3.60) is valid for $x\in B$ and any fixed $y\in D_2$, as the right-hand side is independent of y. Now we simplify the bound on the right-hand side of equation (3.60) slightly to get

$$\begin{aligned}\|f-\mathcal{I}_B^q f\|_{\infty,B} \leq 128eC_{\hat{\gamma}} &\left(1+\frac{1}{\pi}\log(q+1)\right)^3(q+1)\\ &\times\left(1+\frac{\hat{\gamma}\,\mathrm{diam}(B)}{2a}\right)\left(1+\frac{4a}{\hat{\gamma}\,\mathrm{diam}(B)}\right)^{-(q+1)}.\end{aligned} \tag{3.61}$$

We first consider the left-hand side of (3.61) to show that it is the difference between h_0 and a separable function, which is what we're aiming for. Substituting in the definitions of f and $\mathcal{I}_B^q$ yields

$$\begin{aligned}&\|f-\mathcal{I}_B^q f\|_{\infty,B}\\ &\quad=\left\|h_0(k\|x-y\|)-\sum_{j_1,j_2=0}^{q}\hat{f}(x_{1,j_1},x_{2,j_2},y_1,y_2)L_{1,j_1}(x_1)L_{2,j_2}(x_2)\right\|_{\infty,B},\end{aligned}$$

where $\hat{f}(x_1,x_2,y_1,y_2) := h_0(k\|x-y\|)$. Now if $\varphi_{j_1,j_2}(x) := L_{1,j_1}(x_1)L_{2,j_2}(x_2)$, for $j_1,j_2\in\{0,\ldots,q\}$ and if

$$\varphi_j(x) \colon = \begin{cases} \varphi_{j-1,0}(x), & \text{for } j \in \{1,\ldots,q+1\}, \\ \varphi_{j-q-2,1}(x), & \text{for } j \in \{q+2,\ldots,2(q+1)\}, \\ \ldots & \ldots \\ \varphi_{j-(q+1)q-1,q}(x), & \text{for } j \in \{(q+1)q+1,\ldots,(q+1)^2\}, \end{cases}$$

and similarly for

$$\chi_{j_1,j_2}(y) \colon = \hat{f}(x_{1,j_1}, x_{2,j_2}, y_1, y_2), \quad \text{for } j_1, j_2 \in \{0,\ldots,q\},$$

to get χ_j, $j \in \{1,\ldots,(q+1)^2\}$, then the left-hand side of (3.61) becomes

$$\left\| h_0(k\|x-y\|) - \sum_{j=1}^{(q+1)^2} \varphi_j(x)\chi_j(y) \right\|_{\infty,B} =: E, \tag{3.62}$$

i.e. it is the error in the separable approximation derived from the interpolant.

Now we look at the right-hand side of (3.61) to find conditions under which h_0 is arbitrarily close to the separable expansion (3.62). We see that provided the contents of the last bracket > 1, the $-(q+1)$ power dominates the expression as $q \to \infty$. This would mean the right-hand side, and hence the difference between the function and the separable expansion, would converge exponentially to zero as $q \to \infty$. This is what we want, so to ensure the contents of the last bracket are > 1 for some fixed $\hat{\gamma} > 1$ (we shall choose this to equal 2) we introduce the admissibility condition that for some constant $\eta > 0$, $2\eta a > \operatorname{diam}(B)$. Then we obtain

$$E \leq 128eC_{\hat{\gamma}} \left(1 + \frac{1}{\pi}\log(q+1)\right)^3 (q+1)\,(1+2\eta)\,(1+1/\eta)^{-(q+1)}\,. \tag{3.63}$$

Now defining $C(\eta) := 128eC_{\hat{\gamma}}\,(1+2\eta)$ where $\hat{\gamma} = 2$ and $\hat{C}(\eta) := \frac{1}{2}\log(1+1/\eta)$, we can write the right-hand side more succinctly as

$$E \leq C(\eta) \left(1 + \frac{1}{\pi}\log(q+1)\right)^3 (q+1)e^{-2\hat{C}(\eta)(q+1)}.$$

We note that for q sufficiently large the exponential dominates the q and $\log q$ factors, so that when $q \geq q_0(\eta)$, for some $q_0(\eta) > 0$, we have

$$E \leq C(\eta) e^{-\hat{C}(\eta)(q+1)}.$$

Now we wish to find a sufficient condition on q for the right-hand side (and hence the error in the separable approximation) to be $< \varepsilon$. Note

$$C(\eta)\exp(-\hat{C}(\eta)(q+1)) < \varepsilon \quad \text{ if and only if } \quad \exp(\hat{C}(\eta)(q+1)) > \frac{C(\eta)}{\varepsilon}. \tag{3.64}$$

Furthermore

$$\begin{aligned} &\text{if } \varepsilon > C(\eta), \text{ (3.64) holds for all } q > 0, \\ &\text{if } \varepsilon < C(\eta), \text{ (3.64) holds when } q + 1 > \frac{1}{\hat{C}(\eta)} \log \frac{C(\eta)}{\varepsilon}. \end{aligned}$$

Thus, recalling the definition of E,

$$\left\| h_0(k\|x-y\|) - \sum_{j=1}^{(q+1)^2} \varphi_j(x)\chi_j(y) \right\|_{\infty,B} < \varepsilon,$$

when

$$q + 1 \geq \max\left\{ q_0(\eta), \frac{1}{\hat{C}(\eta)} \log\left(\frac{C(\eta)}{\varepsilon} \right) \right\}.$$

Translating this result to be in terms of $p := (q+1)^2$ (and letting $p_0 := (q_0+1)^2$ and $\widehat{C}(\eta) := \hat{C}^2(\eta)$) and recalling the definition of $\|.\|_{\infty,B}$, we obtain the result. □

Remark 3.4.15. *As η increases, the admissibility condition $\eta\text{dist}(D_1, D_2) > \text{diam}(D_1, D_2)$ allows the domains D_1 and D_2 to have bigger diameters and smaller separation between them, i.e. the condition on the domains becomes less restrictive. Obtaining low-rank results for Helmholtz fundamental solutions in general require this admissibility condition to ensure that the domains are well separated for their size (recall the discussion in §1.8 and §3.1). As expected, the low-rank*

result breaks down as $\eta \to \infty$. This can be most easily seen in the preceding Lemma 3.4.14 by examining equation (3.63) *- the bound on the error in the approximation E. Indeed, the right-hand side of this bound tends to infinity as $\eta \to \infty$.*

3.4.5 Finding a bound for $\max_{x\in D_1 y\in D_2} h_0(k\|x-y\|)$

Now we have shown the existence of separable expansions of $\exp(\mathrm{i}k\|x-y\|)$ and $h_0(k\|x-y\|)$. To combine them to find a separable expansion for $H_0(k\|x-y\|)$ need a bound on $\max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)|$ (the reason for this becomes clear in the final assembly of the separable expansion of H_0 in §3.4.6). Since $|h_0(z)| \to 0$ as $z \to \infty$, as in the $n = 0$ case of Lemma 3.4.10, finding a bound for $|z| \geq 1$ is easy, the problem is near zero where the Hankel function has a singularity. However we can show the singularity of h_0 at 0 is of the order of the singularity of log at 0 and hence find a bound for $h_0(z)$ in terms of $\log(z)$. Also, observe that the values of k that we consider (see Definition 3.2.2) imply that we only need to know about the behaviour of $h_0(z)$ in the sector $0 \leq \arg(z) \leq \pi/4$ (where $\arg(z)$ is the argument of z).

Lemma 3.4.16. *Given $R > 0$ there exists $C(R)$ such that*

$$|h_0(z)| \leq C(R)\left(|\log(z)| + 1\right),$$

for all z such that $|z| \leq R$ and $0 \leq \arg(z) \leq \pi/4$ and where log *is the complex logarithm with its branch cut on the negative real axis with base e (usually called* ln*).*

Proof. We first note that by (3.49)

$$|h_0(z)| \leq C_1 \text{ for } \mathrm{Re}(z) \geq 1, \quad 0 \leq \arg(z) \leq \frac{\pi}{4}. \tag{3.65}$$

Also note that $h_0(z)$ is analytic in $0 < Re(z) < \infty$ (recall §3.4.2). This means to prove this result it is sufficient to show that

$$|h_0(z)| \leq C_2(R)(1 + |\log(z)|), \quad \text{for all } z \in \{|z| \leq r,\ 0 \leq \arg(z) \leq \frac{\pi}{4}\}, \tag{3.66}$$

for some $r > 0$. To see this, we note that if $r \geq \sqrt{2}$, (3.65) and (3.66) together

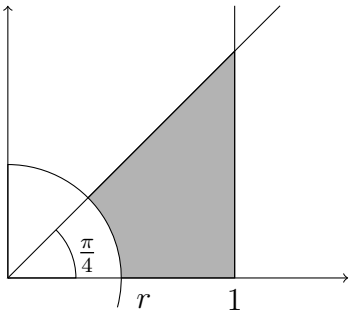


Figure 3-9: The compact set not covered by (3.65) and (3.66) is shaded in grey.

yield the result with $C(R) = C_1 + C_2(R)$. If $r < 1$ we note the set not covered by (3.65) and (3.66) is compact and h_0 is analytic on this set (see for example the grey set in Figure 3-9) and so it is be bounded by some constant $C_3(R)$ and so the result holds with $C(R) = C_1 + C_2(R) + C_3(R)$. To show (3.66) we first note

$$h_0(z) = \exp(-\mathrm{i}z)H_0(z) = \exp(-\mathrm{i}z_R + z_I)H_0(z),$$

so that it is sufficient to show (3.66) holds with the left-hand side replaced by $|H_0(z)|$ (the $\exp(z_I)$ factor can be absorbed in the constant $C_2(R)$). We recall that $H_0^{(1)}(z) = J_0(z) + \mathrm{i}Y_0(z)$ [92, (10.4.3)] and recall the following power series expansions of these Bessel functions:

$$J_0(z) = \sum_{l=0}^{\infty}(-1)^l \frac{\left(\frac{1}{4}z^2\right)^l}{l!\Gamma(l+1)}, \qquad \text{[93, (10.2.2) with } v = 0\text{]}$$

and

$$\begin{aligned} Y_0(z) =& \frac{2}{\pi}\left(\log\left(\frac{1}{2}z\right) + \gamma\right) J_0(z) + \frac{2}{\pi}\left(\frac{\frac{1}{4}z^2}{(1!)^2}\cdots\right. \\ & \left.\cdots - \left(1 + \frac{1}{2}\right)\frac{\left(\frac{1}{4}z^2\right)^2}{(2!)^2} + \left(1 + \frac{1}{2} + \frac{1}{3}\right)\frac{\left(\frac{1}{4}z^2\right)^3}{(3!)^2} - \cdots\right), \quad \text{[93, (10.8.2)]} \end{aligned}$$

where γ is Euler's constant. These expansions then imply that

$$H_0(z) = \frac{2\mathrm{i}}{\pi}\log(z) + \mathcal{O}(1), \quad \text{as } z \to 0,$$

(as $J_0(z) \to 1$, as $z \to 0$ and $\log(z)J_0(z) \to \log(z) + \mathcal{O}(1)$, as $z \to 0$) which gives

(3.66) and hence the result. □

Next we bound the complex $\log(z)$ in terms of the distance of z from the origin.

Lemma 3.4.17. *If* log *is the complex logarithm with its branch cut on the negative real axis, with base e and if* $z = \rho\exp(i\theta)$ *where* $\theta, \rho \in \mathbb{R}$ *then*

$$|\log(z)| \leq |\log(\rho)| + \pi.$$

Proof.

$$\log(z) = \log(\rho\exp(i\theta)) = \log(\rho) + \log(\exp(\mathrm{i}\theta)) = \log(\rho) + \mathrm{i}\theta.$$

Now given where the branch cut is $\theta \in (-\pi, \pi]$ so

$$|\log(z)| = \sqrt{\log^2(\rho) + \theta^2} \leq \sqrt{\log^2(\rho) + \pi^2}.$$

Now $\log^2(\rho) + \pi^2 = \log^2(\rho) + 2|\log(\rho)|\pi + \pi^2 \leq (|\log(\rho)| + \pi)^2$. Combining the last two inequalities gives $|\log(z)| \leq |\log(\rho)| + \pi$, as required. □

Finally in the next lemma we find the bound on $\max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)|$ that we were looking for.

Lemma 3.4.18. *Suppose we have domains and absorption as in Definition 3.2.1 and 3.2.2 and* log *as in Lemma 3.4.17 then there exists* $C > 0$ *such that*

$$\max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)| \leq C\left[\left|\log(4(k_R a)^2 + (k_R d)^2)^{\frac{1}{2}}\right| + 1\right],$$

where C *doesn't depend on* k, d, a *or* b.

Proof. We bound h_0 for small values of $|k|\|x-y\|$, i.e. near the singularity, using Lemma 3.4.16 and Lemma 3.4.17 and for big values of $|k|\|x-y\|$ separately using Lemma 3.4.10. These are then combined to find a bound for all $x \in D_1$ and for all $y \in D_2$.

We first consider the case when $|k|\|x-y\| \leq \sqrt{2}$. By Lemma 3.4.16 with $z := k\|x-y\|$ and $R := \sqrt{2}$

$$|h_0(k\|x-y\|)| \leq C\left(\sqrt{2}\right)\left(|\log(k\|x-y\|)|+1\right),$$

for all $x \in D_1$, $y \in D_2$, such that $|k|\|x-y\| \leq \sqrt{2}$.

Then by Lemma 3.4.17

$$|h_0(k\|x-y\|)| \leq C(\sqrt{2})\left(|\log(|k|\|x-y\|)|+\pi+1\right),$$

for all $x \in D_1$, $y \in D_2$, such that $|k|\|x-y\| \leq \sqrt{2}$.

This implies

$$\max_{\substack{x\in D_1\\ y\in D_2}\\ |k|\|x-y\|\leq\sqrt{2}} |h_0(k\|x-y\|)|$$
$$\leq C\left(\sqrt{2}\right)\left(\max_{\substack{x\in D_1\\ y\in D_2}\\ |k|\|x-y\|\leq\sqrt{2}} |\log(|k|\|x-y\|)|+\pi+1\right).$$

Due to the fact that log is an increasing function we can take the end points of our domain
$\left[\min_{\substack{x\in D_1\\ y\in D_2}}(k\|x-y\|), \sqrt{2}\right]$ and sum the values of the function at these points to find a bound for the function over the whole domain, so that

$$\max_{\substack{x\in D_1\\ y\in D_2}\\ |k|\|x-y\|\leq\sqrt{2}} |h_0(k\|x-y\|)|$$
$$\leq C\left(\sqrt{2}\right)\left[\left|\log\left(\min(k_R^2+k_I^2)^{1/2}\min_{\substack{x\in D_1\\ y\in D_2}}((x_1-y_1)^2+d^2)^{1/2}\right)\right|+\pi+1+\log\sqrt{2}\right]$$
$$\leq C\left(\sqrt{2}\right)\left(\left|\log\left(4(k_Ra)^2+(k_Rd)^2\right)^{1/2}\right|+\pi+1+\log\sqrt{2}\right).$$

Now we consider the case when $|k|\|x-y\| \geq \sqrt{2}$. We wish to apply the $n=0$ case of Lemma to show h_0 is bounded above by a constant, so we must show the

real part of the argument is ≥ 1. Now

$$\sqrt{(k_R^2 + k_I^2)}\|x - y\| \geq \sqrt{2},$$

and the fact that $k_R \geq k_I \geq 0$ and $k_R \geq 1$ gives $k_R^2 \geq k_I^2$, so

$$\sqrt{2k_R^2}\|x - y\| \geq \sqrt{2}, \quad \text{or} \quad k_R\|x - y\| \geq \sqrt{2}/\sqrt{2} = 1.$$

This shows this case satisfies the $z_R \in [1, \infty)$ condition in the $n = 0$ case of Lemma 3.4.10; so by Lemma 3.4.10 we can say

$$|h_0(k\|x - y\|)| \leq C_1,$$

for some $C_1 > 0$, (where C_1 has no x, y, k, d, a or b dependencies),

$$\text{for all } x \in D_1, y \in D_2, \quad \text{such that} \quad |k|\|x - y\| \geq \sqrt{2}.$$

This yields

$$\max_{\substack{x \in D_1 \\ y \in D_2}} \Big|_{|k|\|x-y\| \geq \sqrt{2}} |h_0(k\|x - y\|)| \leq C_1.$$

Finally we combine the two cases to find the bound we want.

$$\begin{aligned}
&\max_{\substack{x \in D_1 \\ y \in D_2}} |h_0(k\|x - y\|)| \\
&\quad \leq \max_{\substack{x \in D_1 \\ y \in D_2}} \Big|_{|k|\|x-y\| \leq \sqrt{2}} |h_0(k\|x - y\|)| + \max_{\substack{x \in D_1 \\ y \in D_2}} \Big|_{|k|\|x-y\| \geq \sqrt{2}} |h_0(k\|x - y\|)| \\
&\quad = C\left(\sqrt{2}\right)\left(\left|\log\left(4(k_R a)^2 + (k_R d)^2\right)^{1/2}\right| + \pi + 1 + \log\sqrt{2}\right) + C_1 \\
&\quad = C\left(\sqrt{2}\right)\left(\left|\log\left(4(k_R a)^2 + (k_R d)^2\right)^{1/2}\right| + \pi + 1 + \log\sqrt{2} + \frac{C_1}{C\left(\sqrt{2}\right)}\right) \\
&\quad \leq C\left[\left|\log\left(4(k_R a)^2 + (k_R d)^2\right)^{1/2}\right| + 1\right],
\end{aligned}$$

where $C := \max\left\{C\left(\sqrt{2}\right), \pi + 1 + \log\sqrt{2} + \frac{C_1}{C(\sqrt{2})}\right\}$, which does not depend on k, d, b or a. □

3.4.6 Final Assembly of New Low-Rank Result and Proof of Related Lemmas

We combine our separable expansions and our bound on the h_0 function to prove Theorem 3.2.3.

Proof of Theorem 3.2.3. To begin with, we write the difference between the Hankel function and a separable expansion in terms of separable expansions of the form we already have from previous results.

$$H_0(k\|x-y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y)$$
$$= \exp\left(\mathrm{i}k\|x-y\|\right)h_0(k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1))\sum_{j=1}^{p}\hat{\phi}_j(x)\hat{\chi}_j(y)$$

for $\hat{\phi}_j(x) := \exp(-\mathrm{i}kx_1)\phi_j(x)$, $\hat{\chi}_j(y) := \exp(\mathrm{i}ky_1)\chi_j(y)$, for all $j \in \{1,...,p\}$. Then

$$H_0(k\|x-y\|) = (\exp(\mathrm{i}k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1)))h_0(k\|x-y\|)$$
$$+ \exp(\mathrm{i}k(x_1-y_1))\left(h_0(k\|x-y\|) - \sum_{j=1}^{p}\hat{\phi}_j(x)\hat{\chi}_j(y)\right).$$

By taking the absolute value and then the maximum over the right-hand side, the triangle inequality yields

$$\left|H_0(k\|x-y\|) - \sum_{j=1}^{p}\phi_j(x)\chi_j(y)\right|$$
$$\leq \underbrace{\max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)|\,|\exp(\mathrm{i}k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1))|}_{\text{Term 1}} \tag{3.67}$$
$$+ \underbrace{\max_{\substack{x\in D_1\\ y\in D_2}} |\exp(\mathrm{i}k(x_1-y_1))|\left|h_0(k\|x-y\|) - \sum_{j=1}^{p}\hat{\phi}_j(x)\hat{\chi}_j(y)\right|}_{\text{Term 2}}.$$

If we show the existence of $\hat{\phi}_j$ and $\hat{\chi}_j$ with Term 1 $\leq \varepsilon/2$ and Term 2 $\leq \varepsilon/2$

then we are done.

Term 1

Let

$$M := \max_{\substack{x\in D_1\\ y\in D_2}} |h_0(k\|x-y\|)| \text{ where } M \text{ is a function of } k,\ a,\ b \text{ and } d.$$

We wish to find the bound

$$M\,|\exp(\mathrm{i}k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1))| \leq \frac{\varepsilon}{2}.$$

This is true provided

$$|\exp(\mathrm{i}k\|x-y\|) - \exp(\mathrm{i}k(x_1-y_1))| \leq \frac{\varepsilon}{2M}.$$

This is the case by Corollary 3.4.7 with $\varepsilon := \varepsilon/2M$ provided

$$\frac{k_R d^2}{a} \leq \exp(k_I a)\frac{\varepsilon}{2M},$$

as all the other conditions are met.

This is tightest when M is biggest, so M can be replaced by something which bounds M above, for which we use Lemma 3.4.18

$$M \leq C\left(\left|\log(4(k_R a)^2 + (k_R d)^2)^{\frac{1}{2}}\right| + 1\right), \tag{3.68}$$

where C doesn't depend on x, y, k, d, a or b, nor our new variables of interest which are not to do with this lemma (namely η and ϵ). Hence provided

$$\frac{k_R d^2}{a}\left(\left|\log(4(k_R a)^2 + (k_R d)^2)^{\frac{1}{2}}\right| + 1\right) \leq C\varepsilon\exp(k_I a), \tag{3.69}$$

(where C has changed from the last line and still doesn't depend on η, ϵ, x, y, k, d, a or b) we have Term $1 \leq \varepsilon/2$.

Term 2

We wish to show our Term 2 $\leq \frac{\varepsilon}{2}$ for some $\hat{\phi}_j(x)$ and $\hat{\chi}_j(y)$. Note

$$\begin{aligned}\max_{\substack{x\in D_1\\ y\in D_2}} |\exp(\mathrm{i}k(x_1-y_1))| &= \max_{\substack{x\in D_1\\ y\in D_2}} |\exp(\mathrm{i}(k_R+\mathrm{i}k_I)(x_1-y_1))| \\ &= \max_{\substack{x\in D_1\\ y\in D_2}} \underbrace{|\exp(\mathrm{i}k_R(x_1-y_1))|}_{=1} |\exp(-k_I(x_1-y_1))| \\ &= \left|\exp\left(-k_I \min_{\substack{x\in D_1\\ y\in D_2}}(x_1-y_1)\right)\right| = |\exp(-k_I 2a)|\,.\end{aligned} \tag{3.70}$$

So Term 2 $\leq \exp(-k_I 2a)\left|h_0(k\,\|x-y\|) - \sum_{j=1}^{p}\hat{\phi}_j(x)\hat{\chi}_j(y)\right|$. It's now sufficient to show Term 2 $\leq \frac{\varepsilon}{2}$ if we show

$$\left|h_0(k\,\|x-y\|) - \sum_{j=1}^{p}\hat{\phi}_j(x)\hat{\chi}_j(y)\right| \leq \frac{\varepsilon}{2\exp(-k_I 2a)}, \tag{3.71}$$

for all $x \in D_1$ and $y \in D_2$. If

$$\frac{\varepsilon}{2\exp(-k_I 2a)} \geq 1,$$

then we can chose p, ϕ_j and χ_j by Lemma 3.4.14 with $\varepsilon := 1$, so that

$$p \geq \max\left\{p_0(\eta), \frac{1}{\widehat{C}(\eta)}\log^2(C(\eta))\right\}.$$

If

$$\frac{\varepsilon}{2\exp(-2k_I a)} \leq 1,$$

then we can choose p, ϕ_j and χ_j by Lemma 3.4.14 with

$$\varepsilon := \frac{\varepsilon}{2\exp(-k_I 2a)},$$

so that

$$p \geq \max\left\{p_0(\eta), \frac{1}{\widehat{C}(\eta)}\log^2\left(\frac{2C(\eta)\exp(-2k_I a)}{\varepsilon}\right)\right\}.$$

So overall we have equation (3.71) with

$$p \geq \max\left\{p_0(\eta), \frac{1}{\widehat{C}(\eta)}\log^2\left(C'(\eta)\max\left\{\frac{\exp(-2k_I a)}{\varepsilon}, 1\right\}\right)\right\},$$

where $C'(\eta) := 2C(\eta)$ so that p_0, $\widehat{C}$ and C' depend only on η. (Thus in particular, $p = \mathcal{O}\left(\log^2\left(\frac{1}{\varepsilon}\right)\right)$ as $\varepsilon \to 0$ for fixed k_I and a.) Therefore Term 2

$$\leq \frac{\varepsilon}{2\exp(-k_I 2a)}\exp(-k_I 2a) = \frac{\varepsilon}{2},$$

with p as above. Thus the result holds, for $C_1 := C$, $C_2(\eta) := \max\{p_0(\eta), 1/\widehat{C}(\eta)\}$ and $C_3(\eta) = C'(\eta)$ and where C_1 doesn't depend upon our parameters of interest and C_2 and C_3 depend only on η. □

Proof of Lemma 3.2.4. We prove this lemma by substituting our conditions into our restrictive condition (3.1). Then the form of the left-hand side of the resulting expression lets us use the fact that $\log(p)p^{\hat{\delta}} \to 0$ as $p \to 0$ for any $\hat{\delta} > 0$, to see the left-hand side $\to 0$ as $p \to 0$. Therefore as $\frac{1}{p} = k_R$, there exists $k_0(\varepsilon)$ such that for all $k_R \geq k_0(\varepsilon)$ the restrictive condition is satisfied.

We note that $h \lesssim a \lesssim 1$ can be covered by considering $a \sim h^\nu$ for $0 \leq \nu \leq 1$. Substituting $a \sim h^\nu$ and $d \sim h$ in (3.1) yields

$$\frac{k_R h^2}{h^\nu}\left[1 + |\log(4(k_R h^\nu)^2 + (k_R 2h)^2)|\right] \lesssim \varepsilon\exp(k_I h^\nu).$$

Rearranging gives

$$k_R h^{2-\nu}\left[1 + |\log 4k_R^2 h^2(h^{2\nu-2}+1)|\right] \lesssim \varepsilon\exp(k_I h^\nu).$$

Using the properties of logs, we then have

$$k_R h^{2-\nu} + k_R h^{2-\nu}|\log(k_R h(h^{2\nu-2}+1)^{1/2})| \lesssim \varepsilon\exp(k_I h^\nu).$$

We substitute in $h = \mathcal{O}(k_R^{-\mu})$ and get

$$k_R k_R^{\mu(\nu-2)} + k_R k_R^{\mu(\nu-2)}|\log(k_R k_R^{-\mu}(k_R^{\mu(2-2\nu)}+1)^{1/2})| \lesssim \varepsilon\exp(k_I k_R^{-\nu\mu}),$$

rearranging we get

$$k_R^{\mu\nu-2\mu+1} + k_R^{\mu\nu-2\mu+1}|\log(k_R^{1-\mu}(k_R^{\mu(2-2\nu)}+1)^{1/2})| \lesssim \varepsilon \exp(k_I k_R^{-\nu\mu}). \tag{3.72}$$

We next show that the power of k_R in both terms in the left-hand side of (3.72), satisfies the following condition: $\mu\nu - 2\mu + 1 \leq -\grave{\delta}$, for some $\grave{\delta} > 0$; then we show that the log in the left-hand side of (3.72) is of the order $\log(k_R)$. From both of these facts about the left-hand side of (3.72), we then know that on the left-hand side of (3.72), the $k_R^{-\grave{\delta}}$ factor dominates the log as $k_R \to \infty$, whilst the right-hand side of (3.72) is $\geq \varepsilon$ as $k_R \to \infty$ and so the condition (3.1) is readily satisfiable for k_R sufficiently large. To show the first fact about the power of k_R on the left-hand side of (3.72), we recall that since $0 \leq \nu < 2 - 1/\mu$ and $1 \leq \mu \leq 2$, we have $-3 \leq \mu\nu - 2\mu + 1 < 0$, so that $k^{-3} \leq k_R^{\mu\nu-2\mu+1} \leq k_R^{-\grave{\delta}}$, for some $\grave{\delta} > 0$. Also $-2 \leq \mu(2-2\nu) \leq 4$, so that

$$1 \leq (k_R^{\mu(2-2\nu)}+1)^{1/2} \leq (k_R^4+1)^{1/2} \leq \sqrt{2}k_R^2,$$

so that

$$k_R^{-1} \leq k_R^{1-\mu}(k_R^{\mu(2-2\nu)}+1)^{1/2} \leq \sqrt{2}k_R^2.$$

Finally $\exp(k_I k_R^{-\nu\mu}) \geq 1$, for all k_R. Thus (3.1) is satisfied for our a, d, h provided

$$k_R^{-\grave{\delta}} + k_R^{-\grave{\delta}}|\log k_R| \lesssim \varepsilon. \tag{3.73}$$

Now since $\log(p)p^{\grave{\delta}} \to 0$ as $p \to 0$, the two terms on the left-hand side of (3.73)$\to 0$ as $k_R \to \infty$, meaning that for a given ε this equation is satisfied for k_R sufficiently large (i.e. there exists $k_0(\varepsilon)$ such that for all $k_R \geq k_0(\varepsilon)$ the restrictive condition (3.1) is satisfied). □

Proof of Lemma 3.2.6. The idea of this proof is to see that under the given conditions the right-hand side of the restrictive condition (3.1) increases at a greater rate than the left-hand side as $k_R \to \infty$ and hence the restrictive condition is satisfied for k_R sufficiently large. We divide through by ε, so that it is sufficient to satisfy

$$\frac{k_R d^2}{a\varepsilon}\left[1 + \left|\log\left(4(k_R a)^2 + (k_R d)^2\right)\right|\right] \lesssim \exp(k_I a). \tag{3.74}$$

We first look at the right-hand side, i.e. the $\exp(k_I a)$ factor and note that

$$k_I a = C\beta k_R^{\delta} h^{\nu} = C\beta k_R^{\delta} k_R^{-\nu\mu},$$

where C is some suitable constant independent of our variables of interest which is allowed to change expression to expression. Hence, since $\nu < \delta/\mu$, $k_I a \to \infty$ as $k_R \to \infty$ and in particular $\exp(k_I a) \sim \exp(k_R^{2\tilde{\delta}})$ as $k_R \to \infty$.

We now look at the left-hand side of (3.74). The argument of the log is bounded as follows

$$k_R^2(a^2 + d^2) \lesssim 4(k_R a)^2 + (k_R d)^2 \lesssim k_R^2(a^2 + d^2).$$

Then the left-hand side of (3.74)

$$\lesssim k_R^{1+\nu\mu} d^2 \log(k_R^2(a^2 + d^2))/\varepsilon.$$

Now since $h \lesssim a^2 + d^2 \lesssim 1$ and since $1/\varepsilon = \mathcal{O}(\exp(k_R^{\tilde{\delta}}))$, the left-hand side of (3.74)

$$\lesssim k_R^{1+\nu\mu} \exp(k_R^{\tilde{\delta}}) \log k_R. \tag{3.75}$$

Now for k_R sufficiently large the left-hand side of (3.74)

$$\lesssim \text{the right-hand side of (3.75)} \lesssim \exp\left(k_R^{3\tilde{\delta}/2}\right).$$

This means the left-hand side of (3.74) grows slower than the right-hand side which $\sim \exp(k_R^{2\tilde{\delta}})$ as $k \to \infty$, and therefore there exists $k_0(\varepsilon)$ such that for all $k_R \geq k_0(\varepsilon)$ the restrictive condition (3.1) is satisfied. □

Proof of Lemma 3.2.8. The idea of the proof is the same as for Lemma 3.2.4, but now we have $\varepsilon = \varepsilon' k_R^{-\mu\nu}$. Hence we follow Lemma 3.2.4 and substitute $a \sim h^{\nu}$, $d \sim h$ and $h = \mathcal{O}(k_R^{-\mu})$ in (3.1) to obtain (3.72), which with the substitution of our new k-dependent $\varepsilon (= \varepsilon' k_R^{-\mu\nu})$ is

$$k_R^{\mu\nu-2\mu+1} + k_R^{\mu\nu-2\mu+1} \left|\log(k_R^{1-\mu}(k_R^{\mu(2-2\nu)} + 1)^{1/2})\right| \leq \varepsilon' k_R^{-\mu\nu} \exp(k_I k_R^{-\nu\mu}).$$

Then dividing through by $k_R^{-\mu\nu}$ we have

$$k_R^{2\mu\nu-2\mu+1}\left(1+\left|\log\left(k_R^{1-\mu}(k_R^{\mu(2-2\nu)}+1)^{1/2}\right)\right|\right)\lesssim \varepsilon' \exp(k_I k_R^{-\mu\nu}).$$

Using the upper bound on ν from the lemma conditions we can bound the power of k_R on the left-hand side:

$$\begin{aligned}\nu &< 1-1/2\mu,\\ \nu-1 &< -1/2\mu,\\ 2\mu(\nu-1) &< -1,\\ 2\mu(\nu-1)+1 &< 0.\end{aligned}$$

Thus we have $2\mu(\nu-1)+1\leq -\delta'$ for some $\delta'>0$, then by the same reasoning as in Lemma 3.2.4 we have that there exists $k_0(\varepsilon')$ such that for all $k_R\geq k_0(\varepsilon')$ the restrictive condition (3.1) is satisfied. □

Proof of Theorem 3.2.9. We follow the proof of Theorem 3.2.3 as far as (3.69) with the exception that we replace ε with ε' where $\varepsilon'=\varepsilon\exp(-k_I a)$. Hence provided

$$\frac{k_R d^2}{a}(|\log(4(k_R a)^2+(k_R d)^2)^{1/2}|+1)\leq C\varepsilon' \exp(k_I a)=C\varepsilon,$$

(where C doesn't depend on our variables of interest) we have Term $1\leq \varepsilon'/2$.

Then we follow the old theorem as far as the end of (3.70) and note

$$\max_{\substack{x\in D_1\\ y\in D_2}}|\exp(-k_I(x_1-y_1))|=\exp(-2ak_I)\leq \exp(-k_I a),$$

so that Term $2\leq (\varepsilon/2)\exp(-k_I a)=(\varepsilon'/2)$ provided

$$\left|h_0(k\|x-y\|)-\sum_{j=1}^{p}\hat{\phi}_j(x)\hat{\chi}_j(y)\right|\leq \varepsilon/2. \tag{3.76}$$

By Lemma 3.4.14 with $\varepsilon=\varepsilon/2$ there exists $\hat{\phi}_j$ and $\hat{\chi}_j$ such that (3.76) holds for

all $x \in D_1$, $y \in D_2$ with

$$p \geq \max\left\{p_0(\eta), \frac{1}{\widehat{C}(\eta)} \log^2\left(\frac{2C(\eta)}{\varepsilon}\right)\right\},$$

where p_0, $\widehat{C}$ and C depend only on η. Thus in particular $p = \mathcal{O}\left(\log^2\left(\frac{1}{\varepsilon}\right)\right)$ as $\varepsilon \to 0$. Thus the result holds, for $C_1 := C$, $C_2(\eta) := \max\{p_0(\eta), 1/\widehat{C}(\eta)\}$ and $C_3(\eta) = 2C(\eta)$ and where C_1 doesn't depend upon our parameters of interest and C_2 and C_3 depend only on η. □

Proof of Lemma 3.2.12. By (3.32)

$$\begin{aligned}
|H_0(k\,\|x-y\|)| &= |\exp(\mathrm{i}k\,\|x-y\|)h_0(k\,\|x-y\|)| \\
&= |\exp(\mathrm{i}k_R\,\|x-y\|)|\,|\exp(-k_I\,\|x-y\|)|\,|h_0(k\,\|x-y\|)| \\
&= |\exp(-k_I\,\|x-y\|)|\,|h_0(k\,\|x-y\|)|\,,
\end{aligned}$$

and by Lemma 3.4.18

$$\begin{aligned}
\max_{\substack{x\in D_1\\ y\in D_2}} |H_0(k\|x-y\|)| &\leq \max_{\substack{x\in D_1\\ y\in D_2}} |\exp(-k_I\|x-y\|)| C\left[\left|\log(4(k_Ra)^2+(k_Rd)^2)^{\frac{1}{2}}\right|+1\right] \\
&= C|\exp(-\beta k_R^\delta \min_{\substack{x\in D_1\\ y\in D_2}} \|x-y\|)|\left[\left|\log(4(k_Ra)^2+(k_Rd)^2)^{\frac{1}{2}}\right|+1\right] \\
&= C|\exp(-2\beta k_R^\delta a)|\left[\left|\log(4(k_Ra)^2+(k_Rd)^2)^{\frac{1}{2}}\right|+1\right] \\
&\to 0 \text{ as } k_R^\delta a \to \infty.
\end{aligned}$$

as $d \leq 1$. □

3.5 Proof of Low-Rank Result for the Green's Function

Proof of Theorem 3.3.3. We check that our domains X and Y are of a size that is valid for Theorem 3.2.3. Then we apply Theorem 3.2.3 to the two Hankel functions in the definition of the Green's function (2.12), combining the two to give us the result.

In checking the domains are valid, we note the Hankel functions have arguments $\|x-y\|$ and $\|x-M(y)\|$ and since the Euclidean norm is translationally invariant only the relative positions of X and Y are important. The translation formulae for shifting X and Y into the positions of D_1 and D_2, in terms of $x \in X$ and $y \in Y$, is as follows: $X = (x_1 + a - x_A, x_2 - (Dm - D)h)$, $Y = (y_1 + a - x_A, y_2 - (Dm - D)h)$. $M(Y)$ is shifted in the same way, i.e. $M(Y) = (y_1 + a - x_A, y_2 + 2(y_2 - (Dm + 1)h) - (Dm - D)h)$. We next find that our translated X and Y and $M(Y)$ fit inside domains of the form in Definition 3.2.1. The size of the domains a and b in Definition 3.2.1 are given in the definitions of a and b in Definition 3.3.1. We define $d_G := d$, where d is as in Definition 3.3.1 and $d_H := d$ where d is as in Definition 3.2.1. Then our translated X and Y and $M(Y)$ fit inside domains of the form in Definition 3.2.1 with $d_H := 2(d_G + h) = 2h(D + 1)$.

We wish to apply Theorem 3.2.3 with $\varepsilon := \varepsilon/2$, this we can now do for two reasons. Firstly, since our domains satisfy the admissibility condition $\eta_1 \mathrm{dist}(X, Y) > \mathrm{diam}(X, Y)$, they also satisfy

$$\eta_2 \mathrm{dist}(D_1, D_2) > \mathrm{diam}(D_1, D_2) \text{ for some } \eta_2 > 0. \tag{3.77}$$

To see this, we observe that

$$\mathrm{diam}(X, Y) = \sqrt{(b-a)^2 + d_G^2},$$

and

$$\mathrm{diam}(D_1, D_2) = \sqrt{(b-a)^2 + (2(d_G + h))^2},$$

so that, combined with $h \leq d$,

$$\mathrm{diam}(D_1, D_2) \leq \sqrt{(b-a)^2 + 16 d_G^2} < 4\sqrt{(b-a)^2 + d_G^2} = 4\mathrm{diam}(X, Y),$$

so that

$$8\eta_1 a = 4\eta_1 \mathrm{dist}(D_1, D_2) = 4\eta_1 \mathrm{dist}(X, Y) > 4\mathrm{diam}(X, Y) > \mathrm{diam}(D_1, D_2),$$

and (3.77) with the admissibility constant $\eta_2 = 4\eta_1$. Secondly (3.18) is the same as (3.1) with $d_H = 2(d_G + h)$ (the factor of $1/2$ in the definition of ε can be

absorbed into C_1, creating a new constant C_4). Thus there exists a p as in (3.2) (an identical formula to (3.19) since the $1/2$ in the definition of ε can be absorbed into $C_3(\eta)$, creating a new constant C_5) and functions $\{\phi_j, \chi_j\}_{j=1}^{p}$ such that

$$\left| H_0(k\|x-y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y) \right| \leq \varepsilon/2,$$

for all $x \in X$, $y \in Y \cup M(Y)$. Taking $R := p$, by the triangle inequality (similarly to Engquist and Ying in [34, proof of Theorem 2.3]) we get that

$$\begin{aligned}
&\left| G^m(x,y) - \sum_{j=1}^{R} \phi_j(x)(\chi_j(y) - \chi_j(M(y))) \right| \\
&\leq \left| H_0(k\|x-y\|) - \sum_{j=1}^{p} \phi_j(x)(\chi_j(y)) \right| + \left| H_0(k\|x-M(y)\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(M(y)) \right| \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&\leq \varepsilon,
\end{aligned}$$

for all $x \in X$, $y \in Y$.

We then define $\Phi_J = \phi_j$ and $\Psi_j = \chi_j(y) - \chi_j(M(y))$, for $j = \{1, \ldots, R\}$ and we have proved the result. □

Chapter 4

Low-Rank Approximation of Schur Complements

4.1 Background on Low-Rank Approximations of Schur Complements

In this chapter we consider $\mathcal{H}$-matrix approximations (see §1.8.3) of Schur Complement matrices (Definition 2.2.5).

Recall from §1.7.2 and §2.2.3 that the approximation of the Schur complement matrices is the key idea of sweeping preconditioners (§1.7, §2). In [34] Engquist and Ying perform numerical experiments with their preconditioner where they approximate the Schur complements using strongly admissible $\mathcal{H}$-matrices. In [46] Gatto and Hesthaven perform numerical experiments with their preconditioner (a similar formulation of the sweeping preconditioner that we consider in §2 with Schur complements like those in Definition 2.2.5) using Hierarchically-Semi-Separable matrices (a variant of $\mathcal{H}$-matrices, see §1.8.3) to approximate the Schur complements.

In [34] Engquist and Ying present theory about approximating the Schur complements using weakly admissible $\mathcal{H}$-matrices (see §2.2.3.2). Despite the importance of approximating the Schur complements in sweeping preconditioners there is not, to the best of the authors' knowledge, any low-rank theory for these Schur complements other than that in [34]. The theory in [34] does not consider the effect of adding absorption, of interest in preconditioners for Helmholtz

problems (see §1.9). Therefore in §4.2 we present our low-rank results for approximating the Schur complements using strongly admissible $\mathcal{H}$-matrices, that considers the effect of adding absorption.

Whilst there are many numerical experiments on the performance of sweeping preconditioners that use $\mathcal{H}$-matrix approximations of Schur complements, there are fewer experiments on considering the $\mathcal{H}$-matrix approximations of Schur complements separately. The single figure [34, Figure 2.3] investigates the approximation rank required to get within a tolerance of 10^{-6} for the weakly admissible off-diagonal blocks of a Schur complement. The analysis of the figure focuses on the fact that the ranks increase greatly for problems with Dirichlet boundary conditions, rather than Sommerfeld boundary conditions. Gatto and Hesthaven compare the ranks of off-diagonal blocks of Schur complements for a Laplace problem and a Helmholtz problem, finding that at the top level of the tree (see §4.2.2), the ranks for required for the Laplace problem are less than in the Helmholtz case [46, Figure 5]. [43] is the most systematic investigation of the ranks of Schur complements, but they are a different formulation of Schur complements than the ones we are considering in Definition 2.2.5. None of these experiments about the Schur complements consider absorption, therefore in §4.2 we present our numerical experiments about the ranks of off-diagonal blocks of Schur complements, which investigate the effect of absorption and other properties.

4.2 Low-Rank Result for Schur Complement Matrices in the Hierarchical Matrix Framework

4.2.1 Idea of New Results

Recall in Chapter 3 we proved that the Green's function $G^m(x, y)$ has a separable expansion when x and y lie in separated domains (see Definition 3.3.1, Theorems 3.3.3 and 3.3.7 and Remarks 3.3.4, 3.3.5, 3.3.6, 3.3.9 and 3.3.10 in §3.3). Since our matrix $\mathbb{G}^m$ is formed by evaluating the G^m at pairs of nodes in Ω_m (see Definition 2.2.7), using the results of §3.3 it is possible to prove the existence of a low-rank approximation to certain off-diagonal blocks of the matrix $\mathbb{G}^m$. The low-rank results are for off-diagonal blocks, because in these blocks the nodes in each pair are separated and therefore fall within the separated domains of the

low-rank results for G^m. We highlight the parallels between our new results and Engquist and Ying's Theorem 2.2.23. Recall that in Theorem 2.2.23 a low-rank result was obtained for an off-diagonal block of $\mathbb{G}^m$ when $D = 1$, using Rokhlin and Martinsson's low-rank result for the Hankel function, Theorem 2.2.22. We use a similar process, but use our low-rank results in the place of Theorem 2.2.22 and explicitly consider as many off-diagonal blocks that appear in a 'standard' $\mathcal{H}$-matrix decomposition as possible, rather than just one off-diagonal block.

The statements of the low-rank results for $\mathbb{G}^m$ appear in §4.2.4. There are three main variants, corresponding to the variants of the low-rank results for G^m.

- Theorem 4.2.29 shows potential benefits due to absorption in the rank, corresponding to Theorem 3.3.3.
- Theorem 4.2.33 shows potential benefits due to absorption in the quality of the approximation for a fixed rank, corresponding to Theorem 3.3.7.
- Theorem 4.2.35 shows potential benefits due to absorption in the same way as Theorem 4.2.33, however it is based on Theorem 3.3.7 where ϵ (the measure of the quality of the approximation) is chosen to be k-dependent. The k-dependence causes a (relatively mild) $\log^2(k_R)$ dependence in the rank for increasing k_R, but a considerable improvement in the quality of the approximation.

The results in this section theoretically justify specific low-rank approximations to $\mathbb{S}_m^{-1}$ and provide pointers on how to develop the approximation methods in practice.

In §4.2.2-4.2.3 we use existing $\mathcal{H}$-matrix theory to deduce the $\mathcal{H}$-matrix partition of the matrix into off-diagonal blocks. Then in §4.2.4 we state the low-rank results for off-diagonal blocks of the matrix $\mathbb{G}^m$.

In §4.3 we numerically verify various properties of the $\mathcal{H}$-Matrix approximations of $\mathbb{G}^m$ (for example we look at the ranks and quality of the approximations for increasing k_R). Due to the relationship between $\mathbb{G}^m$ and $\mathbb{S}_m^{-1}$ discussed in §2.2.3, we also investigate the same properties for $\mathcal{H}$-Matrix approximations of $\mathbb{S}_m^{-1}$.

4.2.2 Construction of $\mathcal{H}$-Matrices

In this section we use existing techniques to devise the 'matrix partition' or 'block decomposition' of 'standard' $\mathcal{H}$-matrices (see §1.8.3 and Figure 2-10). This is by no means original (the theory for 'standard' $\mathcal{H}$-matrices is already well established [56, 57, 61]), but is a necessary set-up for our results about $\mathcal{H}$-matrix approximations of Schur complements in §4.2.4. Recall from §1.8.3 that the blocks to be stored in low-rank form are determined using two cluster-tree structures, combined with an admissibility condition [56, 57, 61].

Rather than describing the process of finding the off-diagonal blocks in the standard language and methods of $\mathcal{H}$-matrices, we do so by going through details of a panel-clustering algorithm. This approach is more accessible and the idea is very similar to the standard $\mathcal{H}$-matrix method when it is applied to matrices that come from the discretisation of PDEs. The panels that we "cluster" are domains (in 1D or 2D) that contain sets of points. (This is a special case of the more general notion of clustering appearing in the general $\mathcal{H}$-Matrix theory.) We then apply the panel-clustering algorithm to our sets of nodes to get the particular $\mathcal{H}$-matrix block structure that can approximate our Schur complement matrices.

4.2.2.1 Cluster Trees and Admissibility Conditions in General

Here we summarise the definition of a generic panel-clustering algorithm, see for example [51] (see [65] and [51, §6.1] and references therein for more details about the panel clustering algorithm).

Definition 4.2.1. ***(Panels Γ and sets of panels $\mathcal{T}$)*** *A panel is an open, connected piece of a domain Γ, for example in a 1D domain Γ, a panel is an open subinterval of Γ. $\mathcal{T}$ is a set of disjoint panels whose closures cover Γ, i.e. $\Gamma = \cup_{\tau\in\mathcal{T}} \bar{\tau}$.*

We now introduce a cluster tree structure. This structure clusters (or groups) panels together, enabling us to easily divide up and then manipulate the set of panels $\mathcal{T}$.

Definition 4.2.2. ***(Cluster tree [51, Definition 6.1])*** *A cluster-tree $\mathbb{T}$ is a tree whose vertices (called 'clusters') consist of unions $\sigma = \cup\{\bar{\tau} : \tau \in \mathcal{T}'\}$ for certain subsets $\mathcal{T}' \subset \mathcal{T}$. These are required to satisfy the following properties.*

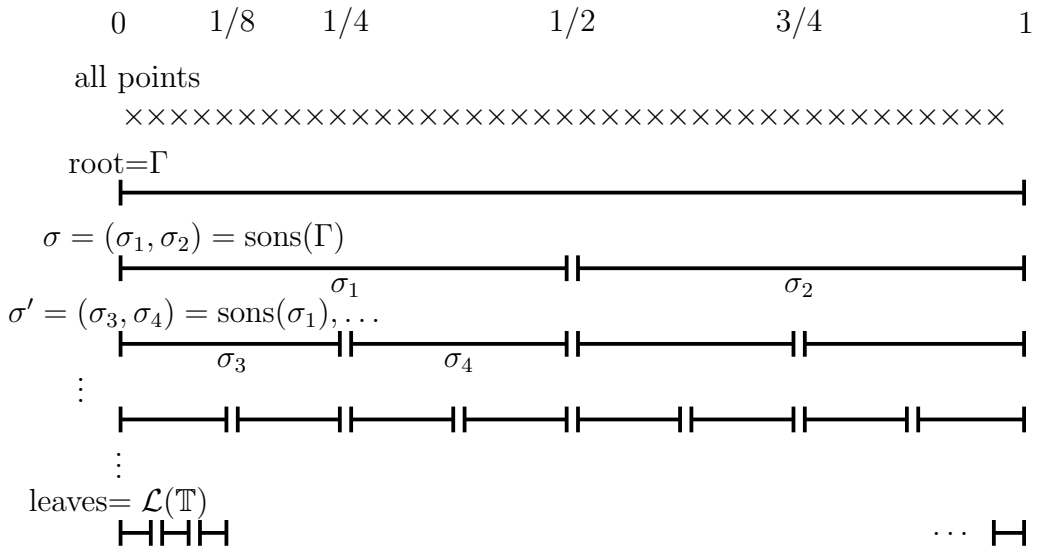


Figure 4-1: Our cluster-tree $\mathbb{T}$ where Γ is an interval is defined by splitting each cluster in half to form the sons on the next level down. Leaves have length $1/2^L$ where $1/2^L = \mathcal{O}(h)$.

(i) $\Gamma = \cup_{\tau \in \mathcal{T}} \bar{\tau}$ *is the root of* $\mathbb{T}$

(ii) $\mathcal{L}(\mathbb{T}) = \mathcal{T}$*, where* $\mathcal{L}(\mathbb{T})$ *denotes the set of leaves of* $\mathbb{T}$

(iii) If $\sigma \in \mathbb{T} \setminus \mathcal{L}(\mathbb{T})$*, there is an associated set of vertices of* $\mathbb{T}$ *(denoted* $\text{sons}(\sigma)$*) which satisfies*

(a) $\sigma = \cup\{\sigma' : \sigma' \in \text{sons}(\sigma)\}$*;*

(b) If σ'*,* $\sigma'' \in \text{sons}(\sigma)$ *and* $\sigma' \neq \sigma''$*, then* σ'*,* σ'' *intersect at most by their boundaries.*

An example of a cluster tree is given in Figure 4-1.

Now we define a second tree whose vertices are pairs of clusters from $\mathbb{T}$, so that we can divide up and manipulate (Γ, Γ).

Definition 4.2.3. ***(Paired Cluster Tree [51, Definition 6.2])***

(i) $(\Gamma, \Gamma) \in \mathbb{T}_2$ *is the root of* $\mathbb{T}_2$

(ii) For $b = (\sigma', \sigma'') \in \mathbb{T}_2$, the set of sons is defined as follows:

$$\operatorname{sons}(b) := \begin{cases} \operatorname{sons}(\sigma') \times \operatorname{sons}(\sigma'') & \text{if } \sigma', \sigma'' \in \mathbb{T} \setminus \mathcal{L}(\mathbb{T}), \\ \{\sigma'\} \times \operatorname{sons}(\sigma'') & \text{if } b \in \mathcal{L}(\mathbb{T}) \times (\mathbb{T} \setminus \mathcal{L}(\mathbb{T})), \\ \operatorname{sons}(\sigma') \times \{\sigma''\} & \text{if } b \in (\mathbb{T} \setminus \mathcal{L}(\mathbb{T})) \times \mathcal{L}(\mathbb{T}), \\ \emptyset & \text{if } b \in \mathcal{L}(\mathbb{T}) \times \mathcal{L}(\mathbb{T}). \end{cases}$$

Definition 4.2.4. ***(Levels)*** *Let the root of a cluster tree be level* 0*. The sons of the root are level* 1*. The sons of these are then level* 2 *and so on.*

Definition 4.2.5. ***(Bottom Level L)*** *The bottom level of a cluster tree, containing all the leaves, is denoted level L.*

Later we run a panel clustering algorithm to form a non-overlapping decomposition of $\mathbb{T}_2$. To give the decomposition the properties we need (i.e. the property that as much of the matrix as possible is covered by off-diagonal blocks that can be approximated in a low-rank way in a $\mathcal{H}$-matrix), we need the concept of admissibility for clusters. We look at two types of admissibility, both commonly found in $\mathcal{H}$-matrix theory.

Definition 4.2.6. ***(Weakly Admissible [51, Definition 6.3])*** *For $\eta > 0$, a pair $(\sigma', \sigma'') \in \mathbb{T}_2$ is called weakly admissible if they are disjoint, i.e.*

$$\sigma' \cap \sigma'' = \emptyset.$$

Definition 4.2.7. ***(Strongly Admissible [51, Definition 6.3])*** *For $\eta > 0$, a pair $(\sigma', \sigma'') \in \mathbb{T}_2$ is called strongly admissible if*

$$\eta \operatorname{dist}(\sigma', \sigma'') \geq \max\{\operatorname{diam} \sigma', \operatorname{diam} \sigma''\},$$

where $\operatorname{dist}(\sigma', \sigma'')$ *is the minimum distance between $\bar{\sigma}'$ and $\bar{\sigma}''$ and* $\operatorname{diam} \sigma'$ *is the maximum distance between any 2 points in $\bar{\sigma}'$.*

Note that strongly admissible is also sometimes called to η-admissible in the literature. A pair being strongly admissible means that their separation is proportional to the diameters of the clusters in the pair, so that they are 'well separated', not simply disjoint.

Then the panel clustering algorithm splits $\Gamma \times \Gamma$ into two sets P_{far} (the far-field) and P_{near} (the near-field) of $\mathbb{T}_2$, using one of the admissibility conditions. The algorithm is as follows.

Algorithm 4.2.8 (**Divide Algorithm [51]**). *First set* $P_{\text{near}} = \emptyset = P_{\text{far}}$ *and then initiate a call* **divide**(Γ, Γ) *to the following recursive procedure:*
procedure divide(σ', σ'')*;*
begin if (σ', σ'') *is [admissible]* **then** $P_{\text{far}} := P_{\text{far}} \cup \{(\sigma', \sigma'')\}$
else if (σ', σ'') *is a leaf* **then** $P_{\text{near}} := P_{\text{near}} \cup \{(\sigma', \sigma'')\}$
else for all $(c', c'') \in \text{sons}(\sigma', \sigma'')$ **do divide**(c', c'')
end.

The union of the sets $P := P_{\text{near}} \cup P_{\text{far}}$ is a non-overlapping covering of $\Gamma \times \Gamma$. This means that P covers $\Gamma \times \Gamma$, i.e. $\cup\{\sigma' \times \sigma'' : (\sigma', \sigma'') \in P\} = \Gamma \times \Gamma$ and all the clusters $\sigma' \times \sigma''$ in this covering intersect by at most their boundaries. Therefore, P is the $\mathcal{H}$-matrix partition or block decomposition when the algorithm is applied to cluster trees resulting from sets of points in Γ.

4.2.3 $\mathcal{H}$-Matrix Decompositions

We use the cluster-tree and block cluster-tree structures to construct a $\mathcal{H}$-matrix decomposition of $\mathbb{G}^m$ (Definition 2.2.7). We look at two cases, $\mathbb{G}^m$ with $D = 1$ and $D > 1$ respectively, that produce $\mathcal{H}$-matrices with fundamentally different structures. (Recall that D is the number of grid rows in Ω_m and therefore the Green's function G^m is evaluated on D grid rows to form $\mathbb{G}^m$, see Definitions 2.2.1, 2.2.2, 2.2.6 and 2.2.7).

4.2.3.1 Case 1: $\mathcal{H}$-Matrix Decomposition of $\mathbb{G}^m$ when $D = 1$

We begin by stating some properties of $\mathbb{G}^m$.

Conditions 4.2.9. *We assume our grid has n interior nodes on any row, where $n = 2^s$ for some integer s and so our $\mathbb{G}^m$ is an $n \times n$ matrix.*

Recall $h = 1/(n+1)$ and that all n nodes lie on the mth row, i.e. within the domain $[mh] \times [0, 1]$.

We go through the process of how to construct a $\mathcal{H}$-matrix in §4.2.2. We define a set with all the points on the row (also see the top row of Figure 4-1).

Definition 4.2.10. ***(Points of row*** m***)*** *On any row m there are n points of the grid in Figure 2-4 as follows*

$$P_m = \{(ih, mh) \quad for \quad i = 1, \ldots, n\}.$$

Next we recall that we want a structured decomposition of $P_m \times P_m$ into pairs of sets of points. We need the sets to be situated within strongly admissible domains of the form in Definition 3.3.1, to make it possible to apply our low-rank results about G^m from §3.3 to these pairs of sets of points. (Note that the pairs of sets of points are the equivalent of the sets X and Y in Theorem 2.2.23: each pair of sets of points defines an off-diagonal block of $\mathbb{G}^m$, for which we obtain the existence of a low-rank approximation.)

To form such a structured decomposition we apply the panel clustering algorithm we described above in §4.2.2. To begin we must define our panels. According to Definition 4.2.1, the panels for a 1D domain are open intervals within an interval Γ: so in this case the panels are open intervals in $\Gamma = [0, 1]$. For $\mathcal{T}$ as in Definition 4.2.1, we choose that each individual panel in $\mathcal{T}$ contains one node. To make the definition of the points and panels clear we precisely define them using some concepts and notation from [34].

Definition 4.2.11. ***(Panels)*** *On any row m there are n panels, they are the following intervals*

$$\mathcal{P}_m = \{((i-1)/n, i/n) \quad for \quad i = 1, \ldots, n\}.$$

Hence we set $\mathcal{T} = \mathcal{P}_m$ ($\mathcal{T}$ as in Definition 4.2.1). The root Γ of a cluster-tree $\mathbb{T}$ (Γ and $\mathbb{T}$ as in Definition 4.2.2) is then the interval $[0, 1]$. Then clusters/vertices σ of $\mathbb{T}$ are intervals containing one or more points, see Figure 4-1. The sons of a particular σ are found by bisecting the interval and thus dividing the sets of points into two groups (recall since $n = 2^s$ this is always possible as we form the tree and we stop when there are only a small number of points in a cluster), see Figure 4-1.

We give notation for each of intervals in Figure 4-1, or equivalently all the clusters in $\mathbb{T}$.

Definition 4.2.12. ***($\mathcal{J}_i^l$)*** *The clusters in $\mathbb{T}$ are denoted $\mathcal{J}_i^l$ where l is the level of*

the cluster (recall Definition 4.2.4) and i which cluster it is on that level, starting with 1 the first cluster on the left. These are as follows:

$$\mathcal{J}_1^0 = \Gamma = [0,1]$$
$$\mathcal{J}_i^l = [(i-1)\text{len}, i\text{len}]$$

for $i \in \{1,\dots,2^l\}$ *and* $l \in \{1,\dots,L\}$*, where* $\text{len} = 1/2^l$ *is the length of an interval on that level.*

Each interval $\mathcal{J}_i^l$ contains a set of points, we define notation for these.

Definition 4.2.13. *(J_i^l) J_i^l for $i \in \{1,\dots,2^l\}$ and $l \in \{1,\dots,L\}$, is defined to be the set of points inside the interval cluster $\mathcal{J}_i^l$, where l is the level of the cluster and i which cluster it is on that level, starting with 1 the first cluster on the left.*

As there is such a close correspondence between J_i^l and $\mathcal{J}_i^l$ that from now on we refer to many properties of J_i^l which are derived from properties of $\mathcal{J}_i^l$, for example we may call J_i^l clusters, referring to the fact that $\mathcal{J}_i^l$ are clusters.

Conditions 4.2.14. *We choose to stop the bisection when the number of points in the sets is $\hat{p} = 2^{s'}$ for some small $s' \in \mathbb{N}$, $s' > 2$.*

Given a cluster-tree $\mathbb{T}$ defined as above, we create our paired cluster-tree $\mathbb{T}_2$ as in Definition 4.2.3. The vertices of $\mathbb{T}_2$ are pairs of intervals, like $\mathcal{J}_i^l \times \mathcal{J}_j^l$, with corresponding pairs of sets of points $J_i^l \times J_j^l$. Recall that, like X and Y in Theorem 2.2.23, these pairs of intervals correspond to blocks of $\mathbb{G}^m$. We define important notation for these blocks as follows.

Definition 4.2.15. ***(Matrix blocks, based on [34, p709],*** $\mathbb{G}_{i,i'}^{m,l}$***)*** *For any clusters $\mathcal{J} = (\mathcal{J}_i^l, \mathcal{J}_{i'}^l) \in \mathbb{T}_2$, the sets of points contained within each interval are $J_i^l, J_{i'}^l$. The corresponding matrix block for the cluster $\mathcal{J}$ is therefore defined to be*

$$\mathbb{G}_{i,i'}^{m,l} := G^m(J_i^l, J_{i'}^l),$$

i.e. the Green's function evaluated at every combination of pairs of points (pairs containing one point from each set J_i^l and $J_{i'}^l$). The columns of the block correspond to points in $J_{i'}^l$ and the rows J_i^l.

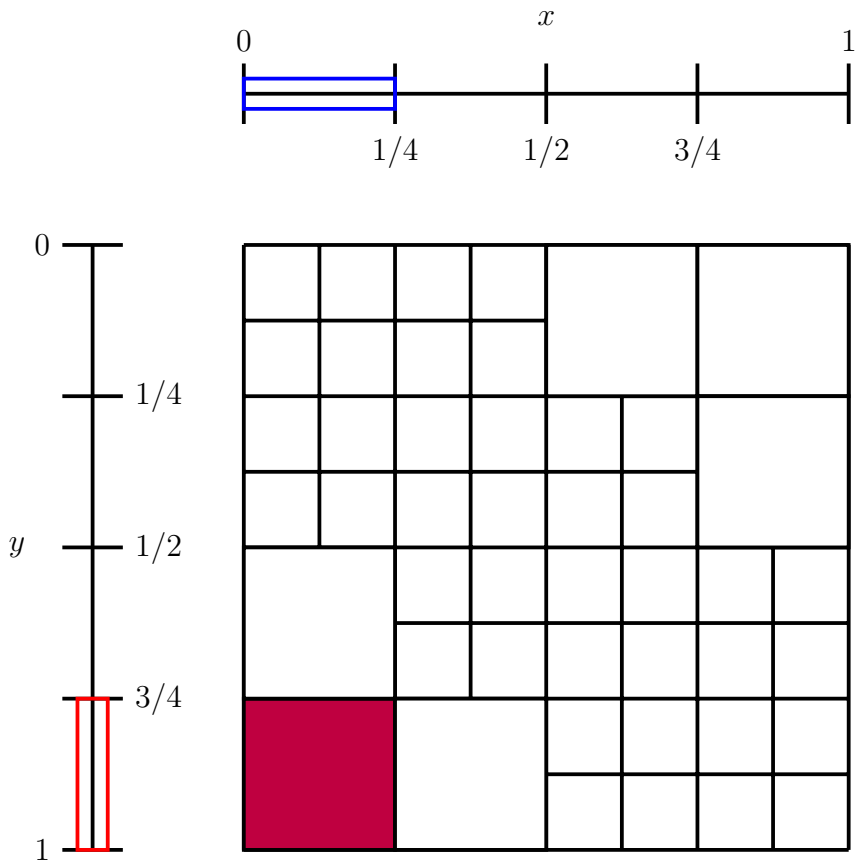


Figure 4-2: Example of how a pair of intervals $[0, 1/4] \times [3/4, 1]$ relates to an off diagonal block of $\mathbb{G}^m \approx \mathbb{S}_m^{-1}$.

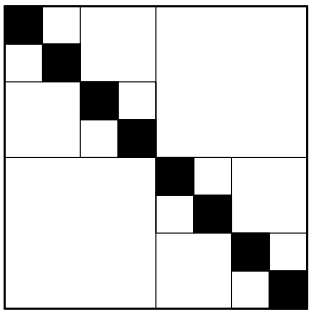


Figure 4-3: A weakly-admissible $\mathcal{H}$-matrix to level 3, the white matrix blocks are those with corresponding clusters in P_{far} and the black matrix blocks are those with corresponding clusters in P_{near}. When approximating the Schur complement matrix $\mathbb{S}_m^{-1}$ using this weakly-admissible $\mathcal{H}$-matrix, white off-diagonal blocks are given a low-rank approximation (as in Figure 1-8) and black blocks are stored densely.

Now we run the divide algorithm, Algorithm 4.2.8, on $\mathbb{T}_2$. As mentioned earlier, this makes a non-overlapping covering/decomposition of $(\mathbb{T}, \mathbb{T}) = \mathbb{T}_2$, where the entries of the decomposition intersect by at most their boundaries. We create the far-field P_{far}, consisting of the admissible clusters corresponding to the off-diagonal blocks that are approximated in a low-rank way in a $\mathcal{H}$-matrix. We also create the near-field P_{near}, consisting of inadmissible leaf-level clusters corresponding to near-diagonal blocks that are small and therefore are stored densely in a $\mathcal{H}$-matrix.

Running the algorithm with different admissibility conditions results in different structures. If we run the algorithm with the weakly admissible condition for clusters (see Definition 4.2.6), we obtain a cluster-tree whose clusters correspond to the matrix structure in Figure 4-3. If we run the algorithm with the strongly admissible condition for clusters with $\eta = 1$ (see Definition 4.2.7), we obtain a cluster-tree whose clusters correspond to the matrix structure in Figure 4-4.

Recall that we want to use our low-rank results for G^m in §3.3, to justify approximating the off-diagonal blocks of $\mathbb{G}^m$ (and hence $\mathbb{S}_m^{-1}$) with low-rank matrix blocks. In order to use these theorems, we need to identity matrix blocks whose corresponding clusters of points lie within domains for which the theorems are valid.

The strongly admissible far-field clusters fit within the theorems' domains. To see this, we go through in more detail how the strongly admissible off-diagonal

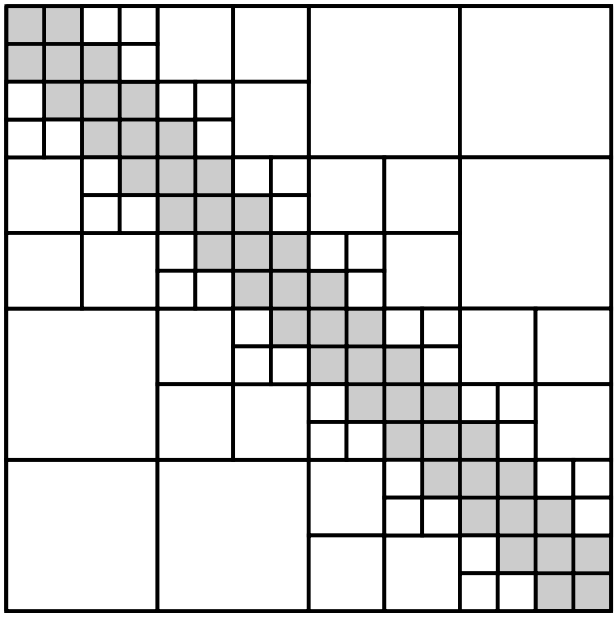


Figure 4-4: A strongly admissible $\mathcal{H}$-matrix to level 4, the white matrix blocks are those with corresponding clusters in P_{far} and the grey matrix blocks are those with corresponding clusters in P_{near}. When approximating the Schur complement matrix $\mathbb{S}_m^{-1}$ using this strongly-admissible $\mathcal{H}$-matrix, white off-diagonal blocks are given a low-rank approximation (as in Figure 1-8) and grey blocks are stored densely.

block structure is created.

Definition 4.2.16. ***(Set of admissible matrix blocks $\mathfrak{G}$,*** *$D = 1$) The set of admissible matrix blocks $\mathfrak{G}$ is defined to be the set of all the matrix blocks $\mathbb{G}^{m,l}_{i,i'}$ whose corresponding clusters $\mathcal{J} := (\mathcal{J}^l_i, \mathcal{J}^l_{i'}) \in \mathbb{T}_2$ appear in the far-field of the strongly admissible decomposition of $\mathbb{T}_2$.*

Note $\mathfrak{G}$ therefore contains matrix blocks for every G^m where $m \in \{1, ..., M\}$.

To get some intuition as to which blocks are in the near and far-field of an strongly admissible $\mathcal{H}$-matrix, we look at the concept of an interaction list, see for example [34, p709].

Definition 4.2.17. ***(Interaction list)*** *The interaction list for a cluster $\mathcal{J}^l_i$ is a list of those clusters on the same level l that are strongly admissible with respect to $\mathcal{J}^l_i$, but whose parents are not strongly admissible to $\mathcal{J}^l_i$'s parent.*

It is easy to see that the sets in $\mathcal{J}^l_i$'s interaction list, also have $\mathcal{J}^l_i$ in their interaction lists, so that inclusion in interaction lists is a reflexive property.

Proposition 4.2.18. *The collection of all the interaction lists is exactly the clusters in P_{far}.*

Proof. By considering the divide algorithm we can say: two clusters $\mathcal{J}^l_i$ and $\mathcal{J}^l_j$ in P_{far} must have parents that aren't strongly admissible, otherwise they will have been included in P_{far} at the previous level. The clusters $\mathcal{J}^l_i$ and $\mathcal{J}^l_j$ in P_{far} must be strongly admissible by the definition of in P_{far}.

An example of the interaction list for a particular set $\mathcal{J}^4_8$ is given in Figure 4-5.

Properties of Domains Arising from $\mathcal{H}$-matrix Decomposition of $\mathbb{G}^m$

We prove a couple of results about the $\mathcal{H}$-matrix block structure needed in the proof of the new low-rank theorems (see statements in §4.2.4 and proofs in §4.2.7).

To apply the low-rank results for the Green's function (Theorems 3.3.3 and 3.3.7) to clusters in P_{far}, corresponding to the far-field matrix blocks, we need to check that the points in $J^l_i \times J^l_{i'}$ lie within domains as in Definition 3.3.1.

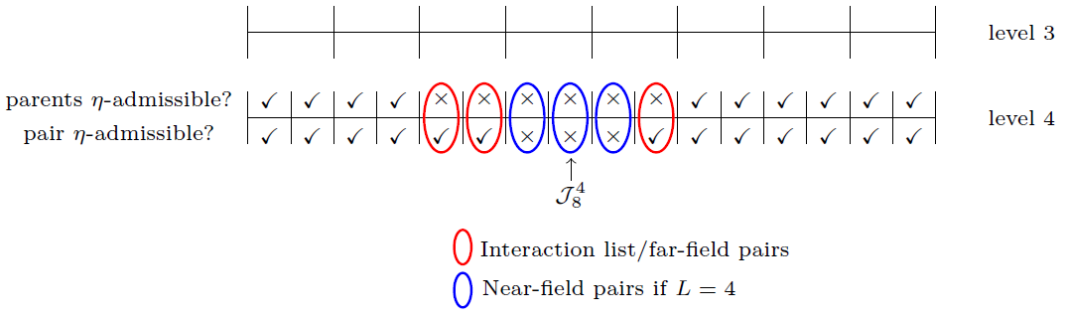


Figure 4-5: For a cluster $\mathcal{J}_8^4$ we see those clusters which pair with it in P_{far} and P_{near}, assuming $L = 4$. If more levels were added the blue, near-field clusters would be further divided. In practice there would be more than 4 levels but this is sufficient to see the pattern. The matrix in Figure 4-4 is also for the strongly admissible block cluster tree to $L = 4$, so that the clusters circled here correspond exactly to particular white and black blocks in Figure 4-4.

Proposition 4.2.19. *All the pairs of clusters in* P_{far} *lie inside domains as in Definition 3.3.1, with* $d = h$ *and* a *and* b *given as follows:*

$$\begin{aligned} a &= \frac{1}{2^{l+1}} \quad \textit{and} \quad b = \frac{3}{2^{l+1}}, \\ a &= \frac{1}{2^{l}} \quad \textit{and} \quad b = \frac{1}{2^{l-1}}, \end{aligned} \tag{4.1}$$

for $l \in \{2, \ldots, L\}$.

Proof. Note that for levels $l = 0$ and $l = 1$ there are no clusters in P_{far}, since on neither of these levels are there any pairs of clusters that satisfy the definition of an admissible pair (Definition 4.2.7).

There is a pattern to the clusters in P_{far} on each level $l \in \{2, \ldots, L\}$: the clusters in each pair in P_{far} are either one cluster apart or two clusters apart, by Proposition 4.2.18 and the pattern visible in Figure 4-5. Therefore, since the clusters are on one row and separated, they lie inside domains as in Definition 3.3.1 for some values of d, a and b. As the sets of points in each cluster are all on one row, a height of $d = h$ is sufficient. The width of a cluster on level l is $1/2^l$, so the separation a of the domains is indeed $1/2^{l+1}$ and $1/2^l$ as in (4.1) for pairs of clusters in P_{far} are either one cluster apart or two clusters apart respectively. Adding the width of a cluster on level l ($1/2^l$) to a then gives the b values $3/2^{l+1}$ and $1/2^{l-1}$ respectively as in (4.1). □

We recall that to be valid domains for the Theorems 3.3.3 and 3.3.7, the domains also need to satisfy the admissibility condition in those theorems.

Proposition 4.2.20. *The clusters in* $\mathrm{P}_{\mathrm{far}}$ *lie inside domains* X *and* Y *that satisfy the admissibility condition* $\eta\mathrm{dist}(X,Y) > \mathrm{diam}(X,Y)$ *for some* $\eta > 0$*, with* $\mathrm{dist}(X,Y)$ *and* $\mathrm{diam}(X,Y)$*, as in Definition 3.3.1.*

Proof. The admissibility condition $\eta\mathrm{dist}(X,Y) > \mathrm{diam}(X,Y)$ is actually already identical to the strongly admissible condition in Definition 4.2.7, since $\mathrm{dist}(\sigma',\sigma'') = a = \mathrm{dist}(X,Y)$ and

$$\mathrm{diam}(\sigma') = \mathrm{diam}(\sigma'') = \sqrt{d^2 + (b-a)^2} = \mathrm{diam}(X,Y). \tag{4.2}$$

□

4.2.3.2 Case 2: $\mathcal{H}$-matrix Decomposition of $\mathbb{G}^m$, $D > 1$

We again use the cluster-tree and block cluster-tree structures in §4.2.2 to construct a $\mathcal{H}$-matrix decomposition of $\mathbb{G}^m$ (Definition 2.2.7). In this case we have $D > 1$ rows of nodes in Ω_m and therefore the Green's function G^m is evaluated on $D > 1$ grid rows to form $\mathbb{G}^m$, see Definitions 2.2.1, 2.2.2, 2.2.6 and 2.2.7.

We construct our $\mathcal{H}$-matrix decomposition of $\mathbb{G}^m$ specifically to apply the low-rank results for the Green's function (Theorems 3.3.3 and 3.3.7) to clusters in $\mathrm{P}_{\mathrm{far}}$, including imposing some properties that ensure the clusters lie within domains as in Definition 3.3.1.

We begin by stating some properties of $\mathbb{G}^m$.

Conditions 4.2.21. *We assume our grid has* n *interior nodes on any row, where* $n = 2^s$ *for some integer* s *and* s *there are* Dn *nodes in* Ω_m *and* $\mathbb{G}^m$ *is a* $Dn \times Dn$ *matrix.*

We go through the process of how to construct an $\mathcal{H}$-matrix in §4.2.2.

The root domain Γ in Definitions 4.2.1 and 4.2.2 is simply Ω_m. The elements of the set $\mathcal{T}$ are panels τ (Definition 4.2.1) that are open rectangular domains as in Figure 4-6.

Definition 4.2.22. ***(Panel width*** p_w ***and panel height*** p_h***)*** *For case* $D > 1$ *with* $\Gamma = \Omega_m$*, for all panels* $\tau \in \mathcal{T}$ *we define the panel width to be* p_w *and the height to be* p_h*.*

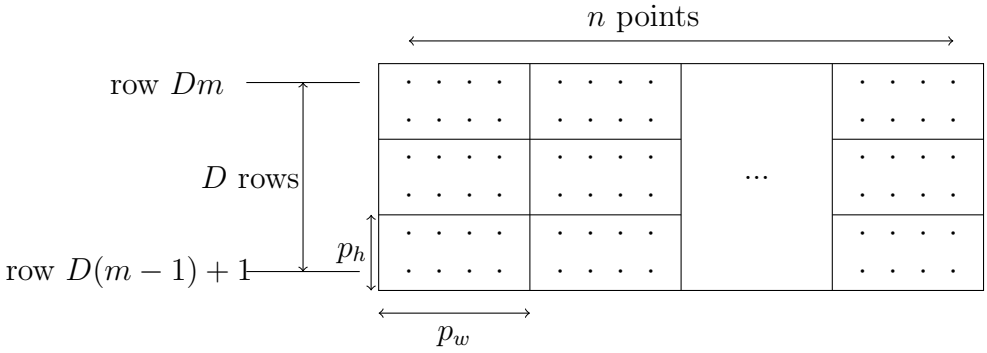


Figure 4-6: The panels $\tau \in \mathcal{T}$ for case $D > 1$. Each leaf panel has height p_h and width p_w.

Using $\mathcal{T}$, we then form the cluster tree $\mathbb{T}$ as in Definition 4.2.2, with entries σ that are unions of panels. Sons are defined by bisecting the non-leaf clusters horizontally (and optionally also subdividing them vertically into equally sized pieces, for example in Figure 4-6 the panels have three vertical subdivisions). It is always possible to bisect horizontally, since $n = 2^s$.

Conditions 4.2.23. *We stop the horizontal bisection when the number of points horizontally in the clusters is $\hat{p} = 2^{s'}$ for some small $s' \in \mathbb{N}$, $s' > 2$. We have at most four vertical subdivisions. d is small enough that*

$$d - 2p_h < \sqrt{p_h^2 + p_w^2}. \tag{4.3}$$

We define a paired cluster tree $\mathbb{T}_2$ as in Definition 4.2.3 using $\mathbb{T}$, which has entries of the form $\sigma \times \sigma'$, i.e. pairs of unions of panels $\sigma \in \mathbb{T}$.

Definition 4.2.24. ***(Levels** l**, separation** a**)** *The levels of the cluster trees $\mathbb{T}$ and $\mathbb{T}_2$ are denoted l, for the roots $l = 0$ and the leaves $l = L$. For each level of $\mathbb{T}_2$ the separation a is defined to be the horizontal width of any $\sigma \in \mathbb{T}$ in any cluster $\sigma \times \sigma'$ in the level, i.e. $a = 1/2^l$.*

We run the divide algorithm (Algorithm 4.2.8) on $\mathbb{T}_2$, using the strong admissibility condition in Definition 4.2.7, with $\eta = 1$. Hence we have a far-field and a near-field non-overlapping decomposition of $\Gamma \times \Gamma$. Each pair of clusters corresponds to a matrix block of $\mathbb{G}^m$ and we give the following notation to these sets of matrix blocks.

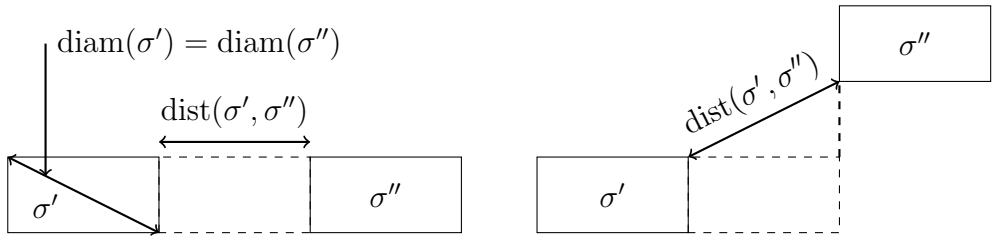


Figure 4-7: Examples of admissible pairs of panels when $D > 1$ that lie inside domains as in Definition 3.3.1. Note especially that the right-hand pair illustrate that the admissible domains do not need to be aligned vertically, providing they are sufficiently separated horizontally.

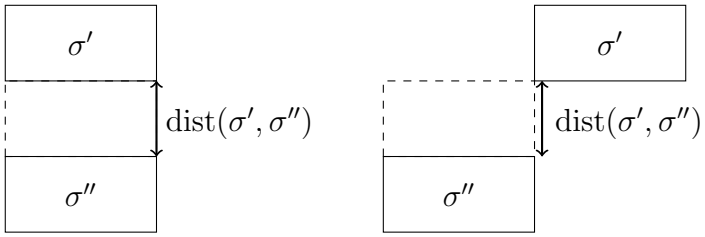


Figure 4-8: Pairs of panels that do not lie in domains like those in Definition 3.3.1.

Definition 4.2.25. ***(Matrix blocks*** $\mathbb{G}^{m,l}_{\sigma,\sigma'}$***)*** *For any clusters* $\sigma \times \sigma' \in \mathbb{T}_2$*, their corresponding sets of points give rise to a block of* $\mathbb{G}$*, defined to be* $\mathbb{G}^{m,l}_{\sigma,\sigma'}$*.*

Definition 4.2.26. ***(Set of admissible matrix blocks*** $\mathfrak{G}$***,*** $D > 1$***)*** *The set of admissible matrix blocks* $\mathfrak{G}$ *is defined to be the set of all admissible matrix blocks* $\mathbb{G}^{m,l}_{\sigma,\sigma'}$ *whose corresponding clusters* $\sigma \times \sigma' \in \mathbb{T}_2$ *appear in the far-field of the strongly admissible decomposition of* $\mathbb{T}_2$*.*

Properties of Domains Arising from $\mathcal{H}$-matrix Decomposition of $\mathbb{G}^{\boldsymbol{m}}$

To apply the low-rank results for the Green's function (Theorems 3.3.3 and 3.3.7) to clusters in $\mathrm{P}_{\mathrm{far}}$, corresponding to the far-field matrix blocks, we need to check that the associated sets of points lie within domains as in Definition 3.3.1. We recall that to be valid domains for the Theorems 3.3.3 and 3.3.7, the domains also need to satisfy the admissibility condition in those theorems.

Proposition 4.2.27. *All the strongly admissible clusters* $\sigma \times \sigma \in \mathrm{P}_{\mathrm{far}}$*, where strongly admissible is according to Definition 4.2.7 with* $\eta = 1$*, and satisfying Assumption 4.2.23, lie inside domains* X *and* Y *as in Definition 3.3.1.* X *and* Y

also satisfy the admissibility condition $\eta\text{dist}(X,Y) > \text{diam}(X,Y)$ *for some* $\mu > 0$*, with* $\text{dist}(X,Y)$ *and* $\text{diam}(X,Y)$*, as in Definition 3.3.1. a in Definition 4.2.24 provides a minimum value of the separation between* X *and* Y*.*

Proof. In order for the strongly admissible clusters to lie inside domains of the form in Definition 3.3.1, all the clusters must have at least one panel in between them horizontally as in Figure 4-7, as opposed to the clusters in Figure 4-8. (Note that the panels in Figure 4-8 are strongly admissible, so that the strongly admissible condition alone is insufficient.) We check this for level $l = L$ as this level has the smallest clusters, that are most likely to fall into the case in Figure 4-8. Since there are at most four vertical subdivisions by Assumption 4.2.23, $\text{dist}(\sigma', \sigma'')$ for panels in Figure 4-8 is at most $d - 2p_h$. Therefore the condition (4.3) in Assumption 4.2.23 ensures that $\text{dist}(\sigma', \sigma'') > \sqrt{p_h^2 + p_w^2}$ is not satisfied for panels of the forms in Figure 4-8.

Since there is at least one panel horizontally separating clusters $\sigma \times \sigma \in \text{P}_{\text{far}}$, the clusters do indeed lie inside domains X and Y as in Definition 3.3.1l. Since the clusters are strongly admissible according to Definition 4.2.7 with $\eta = 1$, X and Y satisfy the admissibility condition $\eta\text{dist}(X,Y) > \text{diam}(X,Y)$ for some $\mu > 0$.

To check that a in Definition 4.2.24 provides a minimum value of the separation between X and Y, we first note that for level L this is already known as there is at least one panel horizontally separating clusters $\sigma \times \sigma \in \text{P}_{\text{far}}$. We can deduce it for the other levels $l \in \{2, ...L-1\}$ as, since their clusters are bigger, they do not fall into the case in Figure 4-8 when the clusters on level L do not. Therefore, they fall into the case in Figure 4-7 and they must have at least one cluster $\sigma \in \mathbb{T}$ of level l in between the cluster pair $\sigma \times \sigma' \in \text{P}_{\text{far}}$ on level l of $\mathbb{T}_2$. □

4.2.4 Statements of New Results

We now state our new low-rank results.

We use the low-rank separable expansion results for the Green's function in §3.3 to find low-rank approximations to the matrix blocks derived in §4.2.3. The full set of conditions for these results are as follows.

Conditions 4.2.28. *Let the model problem be defined as in Definition 1.1.2, with wavenumber including absorption* $k := k_R + \mathrm{i}k_I$. *Assume that there is a discretisation grid on* $[0,1] \times [0,1]$, *with the number of degrees of freedom in one direction* $n = 2^s$ *for some* $s \in N$, $s > 3$ *and grid spacing* $h \sim k_R^{-\mu}$ *for* $1 \leq \mu \leq 2$. *The grid is divided into domains* Ω_m *containing* D *rows as in Definitions 2.2.1 and 2.2.2. Assume that, as in §2.2, in the sweeping preconditioner a series of Schur complements and related half-plane problems arise (see Definitions 2.2.5 and 2.2.18 and Figure 2-7). The Green's functions* G^m *for the half-plane problems are as defined in Definition 2.2.6. The matrices* $\mathbb{G}^m$ *are formed by evaluating the Green's function* G^m *on the grid points in* Ω_m, *as in Definition 2.2.7.*

Let the cluster tree structures and divide algorithm in §4.2.2 be applied to the grid and matrices $\mathbb{G}^m$ *in* Ω_m, *as in §4.2.3.1 and §4.2.3.2. In particular, let Conditions 4.2.9, 4.2.14, 4.2.21and 4.2.23 hold. Let the* $\mathcal{H}$*-matrix decompositions of* $\mathbb{G}^m$ *for* $D = 1$ *and* $D > 1$ *be obtained from the application of the methods in §4.2.3.1 and §4.2.3.2. In particular, let the set* $\mathfrak{G}$ *contain the matrix blocks in the far-field of the strongly admissible decomposition of* $\mathbb{G}^m$, *denoted* $\mathbb{G}^{m,l}_{i,i'}$ *when* $D = 1$ *and* $\mathbb{G}^{m,l}_{\sigma,\sigma'}$ *when* $D > 1$, *as defined in Definitions 4.2.15, 4.2.16, 4.2.25 and 4.2.26, where the index* $l \in \{2, \ldots, L\}$ *corresponds to the level of the corresponding cluster to the block. The value of* a *for any level for* $D = 1$ *is* $1/2^{l+1}$ *from* (4.1) *and for* $D > 1$ a *is as in Definition 4.2.24.*

We now give the new low-rank theorems Theorems 4.2.29, 4.2.33 and 4.2.35, which are the analogues of Theorem 3.3.3, Theorem 3.3.7 and Remark 3.3.10 respectively.

In the first result we make use of the absorption by approximating each block with a an absorption-dependent rank, as in Theorem 3.3.3.

Theorem 4.2.29. ***Low-Rank Result for*** $\mathfrak{G}$ *Let Conditions 4.2.28 hold. Then for all the matrix blocks in* $\mathfrak{G}$, *for a given* $\varepsilon \in (0,1)$, *there exists* $k_0(\varepsilon) > 0$ *such that for all* $k_R \geq k_0(\varepsilon)$, *there exists* R_l, C_1 *and* C_2 *for* $l \in \{2, \ldots, L\}$ *such that*

$$R_l = \left\lceil C_1 \max\left\{1, \log^2\left(C_2 \max\left\{\frac{\exp(-k_I/2^l)}{\varepsilon}, 1\right\}\right)\right\}\right\rceil, \tag{4.4}$$

and

- *when* $D = 1$ *there exists* $\boldsymbol{\Phi_j}, \boldsymbol{\Psi_j} \in \mathbb{R}^{n/2^l}$ *for* $j = \{1, \ldots, R_l\}$, *such that*

$$\left| \left(\mathbb{G}^{m,l}_{i,i'} \right)_{q,q'} - \left(\sum_{j=1}^{R_l} \boldsymbol{\Phi_j} \boldsymbol{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon,$$

  for all $q, q' \in \{1, \ldots, n/2^l\}$,

- *when* $D > 1$, $D = \mathcal{O}(1)$ *there exists* $\boldsymbol{\Phi_j}, \boldsymbol{\Psi_j} \in \mathbb{R}^{n*}$, *where* $n*$ *is the number of points in a level* l *cluster of* $\mathbb{T}$, *for* $j = \{1, \ldots, R_l\}$, *such that*

$$\left| \left(\mathbb{G}^{m,l}_{\sigma,\sigma'} \right)_{q,q'} - \left(\sum_{j=1}^{R_l} \boldsymbol{\Phi_j} \boldsymbol{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon,$$

  for all $q, q' \in \{1, \ldots, n*\}$.

The proof of this result is contained in §4.2.7.

$\mathcal{H}$-matrices are often constructed to have the same rank of approximation in the factorisation of each off-diagonal block, so in the following corollary we find the maximal rank over all the blocks in Theorem 4.2.29; it is sufficient to approximate all the blocks with this one maximal rank. The rank is discussed further in §4.2.5.

Corollary 4.2.30. *Under the same conditions as Theorem 4.2.29, there exists*

$$R = \left\lceil C_1 \max \left\{ 1, \log^2 \left(C_2 \max \left\{ \frac{\exp(-k_I h)}{\varepsilon}, 1 \right\} \right) \right\} \right\rceil, \tag{4.5}$$

and

- *when* $D = 1$ *there exists* $\boldsymbol{\Phi_j}, \boldsymbol{\Psi_j} \in \mathbb{R}^{n/2^l}$ *for* $j = \{1, \ldots, R\}$, *such that*

$$\left| \left(\mathbb{G}^{m,l}_{i,i'} \right)_{q,q'} - \left(\sum_{j=1}^{R} \boldsymbol{\Phi_j} \boldsymbol{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon,$$

  for all $q, q' \in \{1, \ldots, n/2^l\}$,

- *when* $D > 1$, $D = \mathcal{O}(1)$ *there exists* $\boldsymbol{\Phi_j}, \boldsymbol{\Psi_j} \in \mathbb{R}^{n*}$, *where* $n*$ *is the number*

of points in a level l cluster of $\mathbb{T}$, for $j = \{1, \ldots, R\}$, such that

$$\left| \left(\mathbb{G}^{m,l}_{\sigma,\sigma'} \right)_{q,q'} - \left(\sum_{j=1}^{R} \mathbf{\Phi_j} \mathbf{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon,$$

for all $q, q' \in \{1, \ldots, n*\}$.

Proof. This corollary is obtained by taking $C_1 = \max_l\{C_1\}$ and $C_2 = \max_l\{C_2\}$. To find R the maximum rank over all the blocks we must find $\max_l \exp(-k_I/2^l)$. We see that $\max_l \exp(-k_I/2^l) \leq \exp(-k_I h)$, since for the largest value of $l = L$ we have the leaf blocks, whose separation is $\mathcal{O}(h)$ in both cases by Conditions 4.2.14 and 4.2.23 and Propositions 4.2.19 and 4.2.20. □

Corollary 4.2.31. *When $h \sim k_R^{-1}$, for all the matrix blocks in $\mathfrak{G}$ whose corresponding clusters have separation $a \sim h^\nu$ where $0 \leq \nu < 1$, Theorem 4.2.29 and Corollary 4.2.30 hold.*

Proof of Corollary 4.2.31. We use Remark 3.2.5 with $h \sim k_R^{-1}$ (instead of $h \sim k_R^{-\mu}$ for $1 < \mu \leq 2$), when proving Theorem 4.2.29. This introduces the condition $a \sim h^\nu$ with $0 \leq \nu < 1$ from the statement of Lemma 3.2.4, with a, b and d as in Definition 3.3.1. □

Remark 4.2.32. *Note that not all the matrix blocks in $\mathfrak{G}$ satisfy the condition on the separation $a \sim h^\nu$ with $0 \leq \nu < 1$ in Corollary 4.2.31. For example, Conditions 4.2.14 and Proposition 4.2.19 mean that on level L the separation is $\mathcal{O}(h)$, i.e. separation $a \sim h^\nu$ with $\nu = 1$.*

In the second result we make use of the absorption by proving an absorption-dependent improvement to the quality of the approximation, as in Theorem 3.3.7.

Theorem 4.2.33. ***Variant 1 of Low-Rank Result for $\mathfrak{G}$*** *Let Conditions 4.2.28 hold. Then for all matrix blocks in $\mathfrak{G}$, for a given $\varepsilon \in (0, 1)$, there exists $k_0(\varepsilon) > 0$ such that for all $k_R \geq k_0(\varepsilon)$, there exists R, C_1 and C_2 such that*

$$R = \left\lceil C_1 \max\left\{1, \log^2\left(\frac{C_2}{\varepsilon}\right)\right\} \right\rceil, \tag{4.6}$$

and

- *when $D = 1$ there exists $\boldsymbol{\Phi_j}, \boldsymbol{\Psi_j} \in \mathbb{R}^{n/2^l}$ for $j = \{1, \ldots, R\}$, such that*

$$\left| \left(\mathbb{G}^{m,l}_{i,i'} \right)_{q,q'} - \left(\sum_{j=1}^{R} \boldsymbol{\Phi_j} \boldsymbol{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon \exp\left(\frac{-k_I}{2^{l+1}} \right), \tag{4.7}$$

  for all $q, q' \in \{1, \ldots, n/2^l\}$,

- *when $D > 1$, $D = \mathcal{O}(1)$ there exists $\boldsymbol{\Phi_j}, \boldsymbol{\Psi_j} \in \mathbb{R}^{n*}$, where $n*$ is the number of points in a level l cluster of $\mathbb{T}$, for $j = \{1, \ldots, R\}$, such that*

$$\left| \left(\mathbb{G}^{m,l}_{\sigma,\sigma'} \right)_{q,q'} - \left(\sum_{j=1}^{R} \boldsymbol{\Phi_j} \boldsymbol{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon \exp\left(\frac{-k_I}{2^{l+1}} \right),$$

  for all $q, q' \in \{1, \ldots, n*\}$.

The proof of this result is contained in §4.2.7.

Corollary 4.2.34. *When $h \sim k_R^{-1}$, for all the matrix blocks in $\mathfrak{G}$ whose corresponding clusters have separation $a \sim h^\nu$ where $0 \leq \nu < 1$, Theorem 4.2.33 holds.*

Proof of Corollary 4.2.34. We use Remark 3.2.5 with $h \sim k_R^{-1}$ (instead of $h \sim k_R^{-\mu}$ for $1 < \mu \leq 2$), when proving Theorem 4.2.29. This introduces the condition $a \sim h^\nu$ with $0 \leq \nu < 1$ from the statement of Lemma 3.2.4, with a, b and d as in Definition 3.3.1. □

We now create the equivalent of Theorems 4.2.33 where ε depends on k, using Remark 3.3.10.

Theorem 4.2.35. ***Variant 2 of Low-Rank Result for $\mathfrak{G}$*** *Let Conditions 4.2.28 hold. Then for all the matrix blocks in $\mathfrak{G}$ with corresponding clusters appearing in the far-field of the strongly admissible decomposition of $\mathbb{T}_2$ having a separation of at least $a \sim h^\nu$ with $0 \leq \nu < 1 - 1/2\mu$, for a given $\varepsilon' \in (0, 1)$, where ε' is independent of parameters of interest, let $\varepsilon := \varepsilon' k_R^{-\mu\nu}$, then there exists $k_0(\varepsilon') > 0$ such that for all $k_R \geq k_0(\varepsilon')$, there exists R_l, C_1 and C_2 such that*

$$R_l = \left\lceil C_1 \max\left\{ 1, \log^2\left(\frac{C_2 k_R^{\mu\nu}}{\varepsilon'} \right) \right\} \right\rceil, \tag{4.8}$$

where $\nu(l) := \log(a(l))/\log(h)$, *for the separation* $a = 1/2^{l+1}$ *of the corresponding clusters to each block, and*

- *when* $D = 1$ *there exists* $\mathbf{\Phi_j}, \mathbf{\Psi_j} \in \mathbb{R}^{n/2^l}$ *for* $j = \{1, \ldots, R_l\}$, *such that*

$$\left| \left(\mathbb{G}^{m,l}_{i,i'} \right)_{q,q'} - \left(\sum_{j=1}^{R_l} \mathbf{\Phi_j} \mathbf{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon' \exp\left(\frac{-k_I}{2^{l+1}} k_R^{-\mu\nu} \right), \tag{4.9}$$

  for all $q, q' \in \{1, \ldots, n/2^l\}$,

- *when* $D > 1$, $D = \mathcal{O}(1)$ *there exists* $\mathbf{\Phi_j}, \mathbf{\Psi_j} \in \mathbb{R}^{n*}$, *where* $n*$ *is the number of points in a level* l *cluster of* $\mathbb{T}$, *for* $j = \{1, \ldots, R_l\}$, *such that*

$$\left| \left(\mathbb{G}^{m,l}_{\sigma,\sigma'} \right)_{q,q'} - \left(\sum_{j=1}^{R_l} \mathbf{\Phi_j} \mathbf{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon' \exp\left(\frac{-k_I}{2^{l+1}} k_R^{-\mu\nu} \right), \tag{4.10}$$

  for all $q, q' \in \{1, \ldots, n*\}$.

Recall that the range of the k_R dependence is $0 \leq \mu\nu \leq 3/2$.

Remark 4.2.36. *Note that not all the matrix blocks in* $\mathfrak{G}$ *satisfy the condition on the separation* $a \sim h^\nu$ *with* $0 \leq \nu < 1 - 1/2\mu$ *in Theorem 4.2.35. In fact, matrix blocks whose corresponding clusters have separation down to* $\gtrsim h^{1-1/2\mu}$, *or equivalently clusters on levels* $l \lesssim s(1 - 1/2\mu)$, *satisfy Theorem 4.2.35.*

Corollary 4.2.37. *Under the same conditions as Theorem 4.2.35, there exists*

$$R = \left\lceil C_1 \max\left\{ 1, \log^2\left(\frac{C_2 k_R^{\mu - 1/2}}{\varepsilon'} \right) \right\} \right\rceil, \tag{4.11}$$

and

- *when* $D = 1$ *there exists* $\mathbf{\Phi_j}, \mathbf{\Psi_j} \in \mathbb{R}^{n/2^l}$ *for* $j = \{1, \ldots, R\}$, *such that*

$$\left| \left(\mathbb{G}^{m,l}_{i,i'} \right)_{q,q'} - \left(\sum_{j=1}^{R} \mathbf{\Phi_j} \mathbf{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon' \exp\left(\frac{-k_I}{2^{l+1}} k_R^{-\mu\nu} \right),$$

  for all $q, q' \in \{1, \ldots, n/2^l\}$,

- *when $D > 1$, $D = \mathcal{O}(1)$ there exists $\boldsymbol{\Phi_j}, \boldsymbol{\Psi_j} \in \mathbb{R}^{n*}$, where $n*$ is the number of points in a level l cluster of $\mathbb{T}$, for $j = \{1, \ldots, R\}$, such that*

$$\left| \left(\mathbb{G}^{m,l}_{\sigma,\sigma'} \right)_{q,q'} - \left(\sum_{j=1}^{R} \boldsymbol{\Phi_j} \boldsymbol{\Psi_j}^T \right)_{q,q'} \right| \leq \varepsilon' \exp\left(\frac{-k_I}{2^{l+1}} k_R^{-\mu\nu} \right),$$

for all $q, q' \in \{1, \ldots, n*\}$.

Proof. This corollary is obtained by taking $C_1 = \max_l\{C_1\}$ and $C_2 = \max_l\{C_2\}$. To find R the maximum rank over all the blocks we must find $\max_l k_R^{\mu\nu}$. This latter is obtained by taking the largest value of ν in $0 \leq \nu < 1 - 1/2\mu$ (i.e. within the range of ν permitted in Theorem 4.2.35), to obtain $\max_l k_R^{\mu\nu} \leq k_R^{\mu - 1/2}$. □

4.2.5 Discussion of Results

In Theorem 4.2.29 each level/size of blocks potentially has a different rank, in particular a lower rank for more separated or lower level blocks. (In the case $D = 1$, it can be clearly seen that, assuming the $\exp(-k_I/2^l)$ term dominates the inner maximum, the rank in (4.4) decreases as we go away from the diagonal as l (the level of each corresponding cluster) decreases). This phenomenon is, naturally, a result of the similar phenomenon in the analogous results for the Hankel and Green's functions (Theorem 3.2.3 and Theorem 3.3.3); in §3.2.3.1 we see the improvement to the rank required to get within ε due to absorption is related by the factor $\exp(k_I a)$ where a is the separation of the domains. We see more improvements due to absorption for the more separated blocks, since the separation $\mathcal{O}(1/2^{l+1})$ appears in the $\exp(-k_I/2^l)$ factor in the rank (4.4), just as in Chapter 3 we see more improvements due to absorption for the more separated domains.

The fact that some blocks require a lower rank to approximate them in the case of absorption has limited effect on the cost of constructing and applying $\mathcal{H}$-matrix approximations to $\mathbb{G}^m$. The cost of storing or manipulating 'standard' $\mathcal{H}$-matrices of size n, with rank R constant over all the off-diagonal blocks in the low-rank factorised form, is in the form $\mathcal{O}(R^\alpha\ n^\beta \log^\gamma n)$ where $\alpha \in \{1, 2\}$ (see, for example [34, §2.3]). Clearly, if some off-diagonal blocks require a smaller rank, some of the storage/calculational costs are lower, but this would only be

reflected in a lower constant in front of the costing $\mathcal{O}(R^\alpha\ n^\beta \log^\gamma n)$, the overall order is determined by the maximum rank.

Since only the maximum rank makes any difference to the overall order of the costings, $\mathcal{H}$-Matrices are often constructed with the same rank on each block. Therefore we obtained Corollary 4.2.30, which finds the maximal rank (4.5) over all the blocks. The maximum rank (4.5) can be used for every block without affecting the order of the cost. Note that the maximum rank is for the smallest of the off-diagonal blocks, with a separation $a = \mathcal{O}(h)$ and the largest the exponential factor in the rank is for any block is $\exp(-k_I h)$. This means that in this result we see little improvements due to absorption to the costings, since the only place k_I appears is in $k_I h \sim k_I k_R^{-\mu}$ for $1 < \mu \leq 2$, so that as k_R increases this factor actually becomes smaller and the benefits due to absorption disappear as $\exp(-k_I h) \to \exp(0) = 1$ as $k_R \to \infty$.

Theorem 4.2.29 and 4.2.33 prove the existence of a low-rank approximation for all the matrix blocks in $\mathfrak{G}$, with rank R independent of k_R. Since the $\mathcal{H}$-matrices costing $\mathcal{O}(R^\alpha\ n^\beta \log^\gamma n)$ features the rank R, this reduces the expected dependence of the expected costing upon k_R, for k_R sufficiently large. (There is still dependence upon k_R in the costing as n depends on $h \sim k_R^\mu$ for $1 < \mu \leq 2$ (necessary in some cases to combat the pollution effect anyway, see §1.4.2).) This is as opposed to Engquist and Ying's result Theorem 2.2.23, where the rank $R \sim \log(k)$, so that an additional $\log^\alpha(k)$ factor may be expected in the costing $\mathcal{O}(R^\alpha\ n^\beta \log^\gamma n)$. For a full comparison of our results and Engquist and Ying's Theorem 2.2.23, see the next section §4.2.6.

Looking at Theorem 4.2.33, we again see similarities to the analogous Theorem 3.3.7 from which it is produced – here for the same rank of approximation on each block we get an approximation to within ε, but on some blocks this is noticeably improved by the factor $\exp(-k_I/2^{l+1})$. The improvements are definitely seen (although they may be too small to be significant), providing the rank is chosen to be R in (4.6), unlike in Theorem 3.3.3 where improvements are dependent upon whether the exponential term dominates the inner maximum in (4.4).

We also note that, although the improvements due to absorption should always be seen in Theorem 4.2.33, the improvements are once again block dependent – larger, more separated blocks see the biggest improvements. In fact, for increasing k_R, for the specific form of absorption $k_I = \beta k_R^\delta$, with $h \sim k_R^{-\mu}$, for

more separated, lower level blocks that have separation $a = h^\nu = 1/2^{l+1}$ satisfying $0 \leq \nu < 1$ and $\nu < \delta/\mu$, the improvement factor $\exp(k_I/2^{l+1}) \to \infty$, as we saw in §3.2.3.1. For constant k_I, improvements still exist for all blocks, but may be negligible depending upon the relative sizes of k_I and 2^{l+1}.

Looking at Theorem 4.2.35, the rank (4.8) has a $\log^2(k_R^{\mu\nu})$ dependence upon k_R and each level/size of blocks potentially has a different rank, in particular a lower rank for more separated blocks. This dependence upon k_R is inherited directly from the respective results in §3 and is the relatively minor penalty of the much better quality of approximation seen in (4.9) and (4.10), see discussion in §3.2.2. Therefore we obtained Corollary 4.2.37, which finds the maximal rank (4.11) over all the blocks, with the worst k-dependence we see in our results, $\log^2\left(k_R^{\mu-1/2}\right)$.

There is no difference to ranks or quality of approximation between the cases $D = 1$ and $D > 1$, $D = \mathcal{O}(1)$ in any of our theorems. The similarity is down to the second part of the condition, $D = \mathcal{O}(1)$, i.e. that we are considering only quasi-1D domains.

4.2.6 Comparison to Engquist and Ying's Result

We now compare Theorem 2.2.23 of [34] to our Theorems 4.2.29, 4.2.33 and 4.2.35. To assist with the comparison, we reproduce the proof of Theorem 2.2.23 here, with additional explanatory notes and notation adapted to fit in better with the notation used in this thesis.

Before reading the below proof, the reader should recall the statements of Theorems 2.2.22 and 2.2.23 and the domains of the theorems in Figures 2-9 and 4-9.

Proof of Theorem 2.2.23. [34, Theorem 2.3]

Let $d = 2h$ (where d is from Figure 2-9). In order to prove this result we make use of the fact that the Green's function is the sum of two Hankel functions, as in (2.12). Therefore we can apply Theorem 2.2.22 (the result that gives the existence of a low-rank separable expansion for the Hankel function due to Rokhlin and Martinsson [82]) to parts of the sets of points X and Y that satisfy the conditions of Theorem 2.2.22. In particular the sets of points that Theorem 2.2.22 can be applied to must be separated by $2a$ where $a > C(d)|\log\varepsilon|^{\frac{1}{k}}$ (see Figure 2-9).

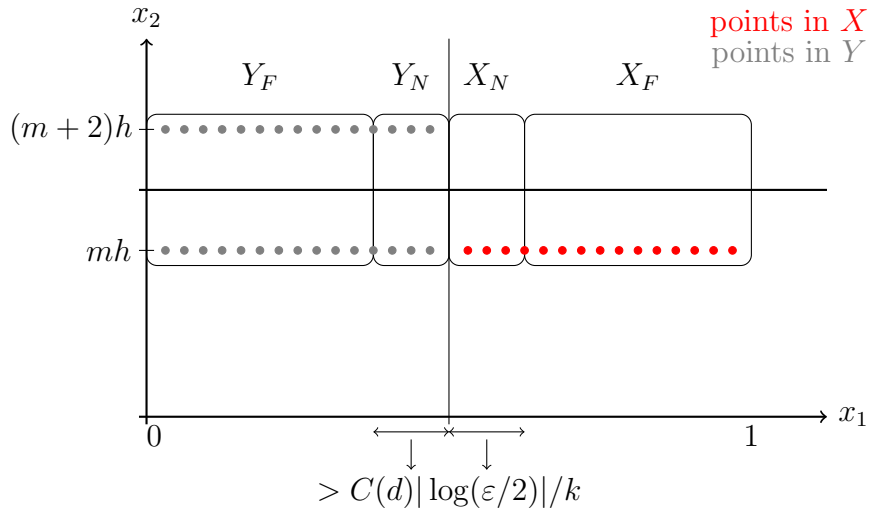


Figure 4-9: The partitions of X and Y.

Hence we define two subsets of X and Y that are separated by $2a$, denoted X_F and Y_F, see Figure 4-9 and equations (2.31) and (2.32). Now Theorem 2.2.22 can be applied to the parts of the matrix whose entries correspond to the points in X_F and Y_F and treat the small remaining parts of the matrix separately to prove the result.

Therefore we have the following block structure for the matrix

$$(G^m(x,y))_{x\in X, y\in Y} = \tag{4.12}$$

$$\begin{pmatrix} (G^m(x,y))_{x\in X_N, y\in Y_N} & (G^m(x,y))_{x\in X_N, y\in Y_F} \\ (G^m(x,y))_{x\in X_F, y\in Y_N} & (G^m(x,y))_{x\in X_F, y\in Y_F} \end{pmatrix}. \tag{4.13}$$

First we consider the $(2,2)$ block of (4.13). We define $M(Y_F)$ as the set of points obtained by reflecting Y_F in the line $x_2 = (m+1)h$ as in the definition of $M(\cdot)$ in Definition 2.2.6). Then $Y_F \cup M(Y_F)$ lies inside the domain

$$\left[0, \frac{1}{2} - C(d)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\frac{1}{k}\right] \times [mh, (m+2)h],$$

and X_F is inside the domain

$$\left[\frac{1}{2} + C(d)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\frac{1}{k}\right] \times [mh, (m+2)h].$$

By construction, these domains satisfy the conditions in Theorem 2.2.22 and so by Theorem 2.2.22 there exists a constant $R \leq \lceil \log(2k) \left|\left(\frac{\varepsilon}{2}\right)\right|^2 \rceil$ and functions $\{\phi_j(x)\}_{1\leq j\leq p}$ and functions $\{\chi_j(y)\}_{1\leq j\leq p}$, such that

$$\left| H_0(k\|x-y\|) - \sum_{j=1}^{p} \phi_j(x)\chi_j(y) \right| \leq \frac{\varepsilon}{2},$$

for $x \in X_F$ and $y \in Y_F \cup M(Y_F)$. Using (2.12) we obtain

$$\left| G^m(x,y) - \sum_{j=1}^{p} \phi_j(x)(\chi_j(y) - \chi_j(M(y)) \right| \leq \varepsilon.$$

Thus we have that this section of the matrix has a low-rank separable approximation of rank R.

We now consider the $(1,1)$, $(1,2)$ and $(2,1)$ blocks of (4.13). X_N and Y_N contain $\mathcal{O}\big(C(2h)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\frac{1}{kh}\big)$ entries (dividing X_N and Y_N's lengths by the grid spacing h). By assumption $kh \sim 1$ so that the hk factor is constant and is included in the constant $C(2h)$. This means that the minimum dimension of the $(1,1)$, $(1,2)$ and $(2,1)$ blocks of (4.13) is $\mathcal{O}\big(C(2h)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\big)$. Hence $\mathcal{O}\big(C(2h)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\big)$ is an upper bound for the ranks of these blocks of (4.13) (and therefore also an upper bound on the matrix rank required to approximate these blocks of (4.13) perfectly).

Since we now know the rank of approximation required to get each entry of each block of $(G^m(x,y))_{x\in X, y\in Y}$ in (4.13) to within ε, we take the maximum of these ranks to obtain the rank we need to approximate the whole matrix to within ε as follows:

$$p = \mathcal{O}\left(3C(2h)\left|\log\left(\frac{\varepsilon}{2}\right)\right| + \log(k)\left|\log\left(\frac{\varepsilon}{2}\right)\right|^2\right) \leq \mathcal{O}\left(C(2h)\log(k)\left|\log\left(\frac{\varepsilon}{2}\right)\right|^2\right), \tag{4.14}$$

for some $C(2h)$. Then there exists $\{\alpha_r(x)\}_{1\leq r\leq p}$ for $x \in X$ and $\{\beta_r(y)\}_{1\leq r\leq p}$ for $y \in Y$ such that

$$\left| G^m(x,y) - \sum_{r=1}^{p} \alpha_r(x)\beta_r(y) \right| \leq \varepsilon \quad \text{for } x \in X,\ y \in Y.$$

□

Remark 4.2.38. *Overall the dependence upon k of the rank p as $k \to \infty$ is $\mathcal{O}(\log(k))$, see the right-hand side of the rank (4.14). To see this involves a little work due to the subtlety that there is a dependence upon k in the factor $C(2h)$, since $h = \mathcal{O}\left(\frac{1}{k}\right)$. However, we deduce that dependence upon k of the factor $C(2h)$ is only potentially beneficial to the rank as $k \to \infty$. To deduce this, we recall that the $2h$ in this factor comes from the height of the domains in Theorem 2.2.22. Therefore the factor $C(2h)$ is referring to the fact that the domain size is getting smaller as $k \to \infty$. Since the same or lower rank is needed to form the approximation on smaller domains, the factor $C(2h)$ will indeed be only potentially be beneficial for increasing k. Since we do not know the dependence upon k of the benefit of the $C(2h)$ factor, we remove it from consideration and we have only the $\mathcal{O}(\log(k))$ dependence left.*

Note there are a number of differences between this Theorem 2.2.23 and our Theorems 4.2.29, 4.2.33 and 4.2.35.

1) Our theorems are valid for complex k and show improvements when absorption is added, Theorem 2.2.23 is only valid for real k.

2) Our theorems consider the case $D > 1$, Theorem 2.2.23 does not.

3) Both our theorems and Theorem 2.2.23 have identical ε-dependence in the rank $\mathcal{O}(\log^2(1/\varepsilon))$, see Theorem 2.2.23's expression for the rank p and (4.4), (4.6) and (4.8), as inherited from the respective results in §3, see discussion in §3.2.4.

4) Our Theorems 4.2.29 and 4.2.33 have no k-dependence in the ranks (see (4.4) and (4.6)), whereas Theorem 2.2.23 has a $\log(k)$ k-dependence. This is an advantage for our results, see comment in §4.2.5. Theorem 4.2.35 has a $\log^2(k^{\mu\nu})$ k-dependence, which is potentially worse than Theorem 2.2.23's $\log(k)$ k-dependence, depending upon where in the range $0 \leq \mu\nu \leq 3/2$ the value for the matrix block lies. Note that, the less separated the corresponding clusters are, the less the k-dependence (for example, consider $a = \mathcal{O}(h)$ i.e. $\nu = 0$, where there is no k-dependence). This is a direct consequence of the $k^{-\mu\nu}$ factor included in ε, which makes the quality of the approximation better in (4.9) and (4.10). The k-dependence is also directly inherited from the respective results in §3, see discussion in §3.2.4.

5) Theorem 2.2.23 only explicitly considers one of the largest off-diagonal blocks of $\mathbb{G}^m$ for a weakly admissible $\mathcal{H}$-matrix (though it can readily be scaled to other blocks), when $h \sim k^{-1}$. Our result is valid for a strongly admissible $\mathcal{H}$-matrix and explicitly describes which blocks can be covered when $h \sim k^{-\mu}$, for $1 \leq \mu \leq 2$. (In some cases all the blocks can be covered, in others not, for example when $h \sim k^{-1}$ in Corollary 4.2.31).

   Note that it is harder to prove the result for a weakly admissible $\mathcal{H}$-matrix than a strongly admissible $\mathcal{H}$-matrix, since the blocks are less well-separated in the weak case (for example, compare Figures 4-3 and 4-4). Hence Theorem 2.2.23 is better in this regard; in practice as Engquist and Ying use strongly admissible $\mathcal{H}$-matrices in their numerical experiments, so this is not a disadvantage for our theorems in practice.

6) Theorem 2.2.23 only considers $h \sim k^{-1}$, whereas our theorems consider a wider range of h values, i.e. $h \sim k_R^{-\mu}$ for $1 \leq \mu \leq 2$. Thus grids which are fine enough to counter the pollution effect are considered. It would also be difficult to adapt Theorem 2.2.23 to this wider range in its present form.

   To see that this would be difficult, we look at the proof the Theorem 2.2.23. The proof involves separating the matrix into blocks and finding the rank of the $(1,1)$, $(1,2)$ and $(2,1)$ matrix blocks of (4.13) using their size, and the size depends on h. Therefore if $h \sim k_R^{-\mu}$ for $1 \leq \mu \leq 2$, the rank (4.14) becomes

$$\begin{aligned} p &= 3C(2h)\left|\log\left(\frac{\varepsilon}{2}\right)\right|\frac{1}{kh^{-\mu}} + \log(k)\left|\log\left(\frac{\varepsilon}{2}\right)\right|^2 \\ &\lesssim C(2h)\log(k)k^{\mu-1}\left|\log\left(\frac{\varepsilon}{2}\right)\right|^2, \end{aligned}$$

   so that the rank grows as a power of k as $k \to \infty$ for $1 < \mu \leq 2$.

   One way the growth in k of the rank can easily be avoided, is to change the method of proof to the one we use, i.e. to remove the need to approximate the $(1,1)$, $(1,2)$ and $(2,1)$ matrix blocks of (4.13) separately, by considering domains from the strongly admissible $\mathcal{H}$-matrices (instead of weakly admissible $\mathcal{H}$-matrices). Using this alternative method of the proof would remove the restricted range of h with respect to k.

4.2.7 Proof of Low-Rank Results for $\mathbb{G}^m$

Proof of Theorem 4.2.29. As we said previously, the idea of this proof is to apply Theorem 3.3.3, via the Remark 3.3.4, to domains containing the strongly admissible, far-field sets of points that each correspond to a matrix block in $\mathfrak{G}$. Then we evaluate the resulting separable expansion for the Green's function on the sets of points, obtaining two sets of R vectors that are a low-rank approximation to the matrix block in $\mathfrak{G}$, as desired.

Case $D = 1$

In this case, we know the sets of points explicitly: various pairs $J_i^l \times J_{i'}^l$, for $i \in \{2, \ldots, 2^l\}$ and $l \in \{1, \ldots, L\}$ and their corresponding matrix blocks $\mathbb{G}_{i,i'}^{m,l}$ (see Definitions 4.2.13 and 4.2.15).

First, note that we have the same boundary value problem and Green's function as in Theorem 3.3.3. To use Remark 3.3.4 we must check that the points in J_i^l and $J_{i'}^l$ lie within domains X and Y of the form in Definition 3.3.1 that Theorem 3.3.3 and Remark 3.3.4 are valid for. Hence we look at the sets of points and define domains they lie in, finally checking each condition these must satisfy in turn. Proposition 4.2.19 allows us to define domains X and Y in the obvious way: by placing the sets of points J_i^l and $J_{i'}^l$ in boxes of sizes d, a and b which go up as far as $x_2 = mh$ as in Figure 3-7. We see that $d = h$ and (4.1) give us sufficient d, a and b values for all the different sets of points on any of our levels $l \in \{2, \ldots, L\}$.

We must now check that the values of d, a and b satisfy all the conditions in Definition 3.2.1, Theorem 3.3.3 and Remark 3.3.4. We have already seen they satisfy the admissibility condition $\eta \text{dist}\{X, Y\} > \text{diam}\{X, Y\}$ from Theorem 3.3.3 (this is by construction since we used the strong or η-admissibility condition in the construction of the $\mathcal{H}$-Matrix, see Proposition 4.2.20). We then check the conditions in Remark 3.3.4 with $\varepsilon := \varepsilon/2$, note that many of the conditions required are those in Lemma 3.2.4, with a, d and b as in Definition 3.3.1. Clearly $d = h$ satisfies $d \sim h$. We note that it is sufficient to consider $h \lesssim a \lesssim 1$ to cover all values of μ in the range of the condition on a in Remark 3.3.4. Our smallest a values occur when $l = L$, so that $a = 1/2^{L+1}$ and $1/2^L$ from (4.1). Recall from Conditions 4.2.14 that the minimum number of points in a leaf cluster of points is $\hat{p} = 2^{s'}$ for some small $s' \in \mathbb{N}$, $s' > 2$. So, $1/2^L = h\hat{p}$ for some $\hat{p} > 8$, so that

$1/2^{L+1} \gtrsim h$, and hence all of our domains have $a \gtrsim h$. Our domains are clearly $\lesssim 1$ as well, so by Remark 3.3.4 we have that for k_R sufficiently large we have satisfied (3.18) in Theorem 3.3.3.

Case $D > 1$

In this case we do not know the sets of points or the off-diagonal matrix blocks in $\mathfrak{G}$ explicitly; however, we can show they satisfy the conditions we require by our construction of them in §4.2.3.2. First, again we have the same boundary value problem and Green's function as in Theorem 3.3.3. We know the sets of points will lie within domains X and Y of the form in Definition 3.3.1 and that X and Y satisfy the condition $\eta a > \operatorname{diam}(X, Y)$ in Theorem 3.3.3, by construction, see Proposition 4.2.27.

We then check the conditions in Remark 3.3.4, with $\varepsilon := \varepsilon/2$. Clearly, since $D = \mathcal{O}(1)$, $d = Dh$ satisfies $d \sim h$. As in the case $D = 1$, that all the values of $a \lesssim 1$ is clear, so that it remains to check that $a \gtrsim h$. This follows similarly to the case $D = 1$. All the admissible domains are separated by at least the horizontal width of a panel p_w. Recall from Conditions 4.2.23 that $p_w = \hat{p}h$, where $\hat{p} \geq 8$, so a is bounded below by $p_w = \hat{p}h = 1/2^{L+1}$, so $a \gtrsim h$.

Both cases

Thus for k_R sufficiently large our domains satisfy all the conditions and hence we can apply Theorem 3.3.3 via Remark 3.3.4 to get that there exists an R_l as required in (3.19) with $a = 1/2^{l+1}$ (see (4.1) for $D = 1$ and Proposition 4.2.27 for $D > 1$) and functions $\{\Phi_j, \Psi_j\}_{j=1}^{R_l}$ such that

$$\left| G(x, y) - \sum_{j=1}^{R_l} \Phi_j(x)\Psi_j(y) \right| \leq \varepsilon,$$

for all $x \in X$ and $y \in Y$. Note each R_l may take a slightly different form as a varies from level to level in (3.19).

Obviously since the sets of points lie in these domains X and Y respectively by construction, we have proved the result. To see this explicitly in case $D = 1$, note we can define $(\mathbf{\Phi}_j)_q := \Phi_j(x_q)$ where x_q is the qth point in J_i^l and $(\mathbf{\Psi}_j)_{q'} := \Psi_j(y_{q'})$ where $y_{q'}$ is the q'th point in $J_{i'}^l$. The case $D > 1$ is similar, but cannot be written explicitly as we do not have explicit notation for all the sets of points/clusters. □

Proof of Theorem 4.2.33. Note that we wish to apply Theorem 3.3.7 via Remark

3.3.9 in the same way as we proved Theorem 4.2.29 via Theorem 3.3.3 and Remark 3.3.4. Since the conditions in Theorem 3.3.7 and Remark 3.3.9 are very similar to the conditions in Theorem 3.3.3 and Remark 3.3.4 and the conditions in Theorem 4.2.33 are very similar to the conditions Theorem 4.2.29, very little remains to be done. One difference is that (3.24) doesn't have the $\exp(k_I a)$ factor on the right-hand side which (3.7) does, but this difference is negated by the remarks (i.e. they give us (3.1) and (3.24) are satisfied for k_R sufficiently large under identical conditions) and so nothing needs to be done about this. Another difference is the equation for the rank R is different, but this is just carried straight through to the statement of Theorem 4.2.33, so that (3.25) is identical to (4.6). Finally the quality of the approximation is $\varepsilon' = \varepsilon \exp(-k_I a)$ instead. On level l this means it is sufficient to say $\varepsilon' = \varepsilon \exp(-k_I/2^{l+1})$ by taking the minimum possible separation value a from (4.1) (Case $D = 1$) and Proposition 4.2.27 (Case $D > 1$) on each level and so we have proved the result. □

Proof of Theorem 4.2.35. Note that the origins of Theorem 4.2.33 are from Theorem 3.3.7 via Remark 3.3.9 and we can get the proof of Theorem 4.2.35 in the same way by applying Remark 3.3.10 instead. The only differences are an additional restriction on the separation of the domains (the $a \sim h^\nu$ with $0 \leq \nu < 1 - 1/2\mu$) and the k-dependence in ε. The former is incorporated into the conditions in the statement of Theorem 4.2.35. The latter is transferred to the rank and separable expansion in the statement of Theorem 4.2.35. In order to have the same rank on each block we take the maximum over all the levels which is when $\nu < 1 - 1/2\mu$, covered by $\nu = 1 - 1/2\mu$. □

4.3 Numerical Verification of Low-Rank Results for $\mathbb{G}^m$

In this section we perform some experiments to numerically verify aspects of our low-rank results for $\mathbb{G}^m$ in §4.2. We look at properties of blocks of four types of matrices.

Type A The matrix $\mathbb{G}^m$ defined in Definition 2.2.7 as the direct evaluation of the Green's function G^m, as in Definition 2.2.6.

Type B An inverse Schur complement block $\mathbb{S}_m^{-1}$, as defined in Definition 2.2.5, with $D = 1$, for a FEM discretisation, as described in §1.4 and §2.1.3, of the homogeneous Helmholtz problem with PML on three sides, see §2.1.1.

Type C Type B but with heterogeneity of two different wavespeeds (**CH3** in §4.3.1.1).

Type D Type B with heterogeneous wavespeed drawn from Marmousi model (**CM** in §4.3.1.1).

We look at all four types to see how well low-rank properties of the Green's function (as seen in the theory in §4.2 and numerics in this section of the corresponding Type A matrix blocks) translate to both to directly equivalent blocks of Schur complement matrices that might actually be found in a sweeping preconditioner (Type B) and similar blocks but with variable wavespeed like those used in practice (Types C and D), for which there is no matching analysis.

We break down the description of these experiments into the next three subsections on

1) §4.3.1 description of how the matrix blocks are formed,

2) §4.3.2 recap of the SVD and ε-rank,

3) §4.3.3-4.3.6 the results,

respectively.

4.3.1 Formation of Matrix Blocks

The formation of **Type A** is given by Definition 2.2.7.

Type B We construct the FEM discretisation, as described in §1.4 and §2.1.3, of the homogeneous Helmholtz problem with PML on three sides, see §2.1.1. We remove the PML from the top of the grid and use a Dirichlet condition on the top of the grid instead, so that we are solving a half-plane problem. We use the inverse Schur complement block $\mathbb{S}_M^{-1}$, as defined in Definition 2.2.5, with $D = 1$. (To find $\mathbb{S}_M^{-1}$, we calculate A^{-1} and select the (M, M)th block, since the discretisation matrix corresponds to that in the Mth half-plane problem in Definition 2.2.18, and by Proposition 2.2.14 $\mathbb{S}_M^{-1}$ is then simply the (M, M)th block of A^{-1}.)

On the other three sides we use PMLs with parameters η and C (see description in §2.1.1) as follows

$$\eta = \min\{0.09 \min\{1, 2\pi/k\}\},$$

and

$$C = 2/\mu.$$

The off-diagonal blocks selected (unless specified otherwise for a particular experiment) are blocks with rows/columns of nodes in sets X/Y where

$$X = \{[jh, 1]^T,\, j \in \{1, \ldots, \text{npts}\}\},$$

and

$$Y = \{[1 - jh, 1]^T,\, j \in \{1, \ldots, \text{npts}\}\},$$

where $\text{npts} = \text{round}(n - 2 - \text{rpts})$, and where $\text{rpts} = 1$ if $a = h$ and is chosen to be the smallest so that the separation of the clusters for the off-diagonal block is at least $2a$ in any other case, chosen to ensure so that an equal distance from the centre is removed (i.e. rpts is even if n is even and odd if n is odd).

Type C and D These are formulated in the same way as Type B but with variable wavespeeds $c(x)$ **CH3** and **CM** respectively.

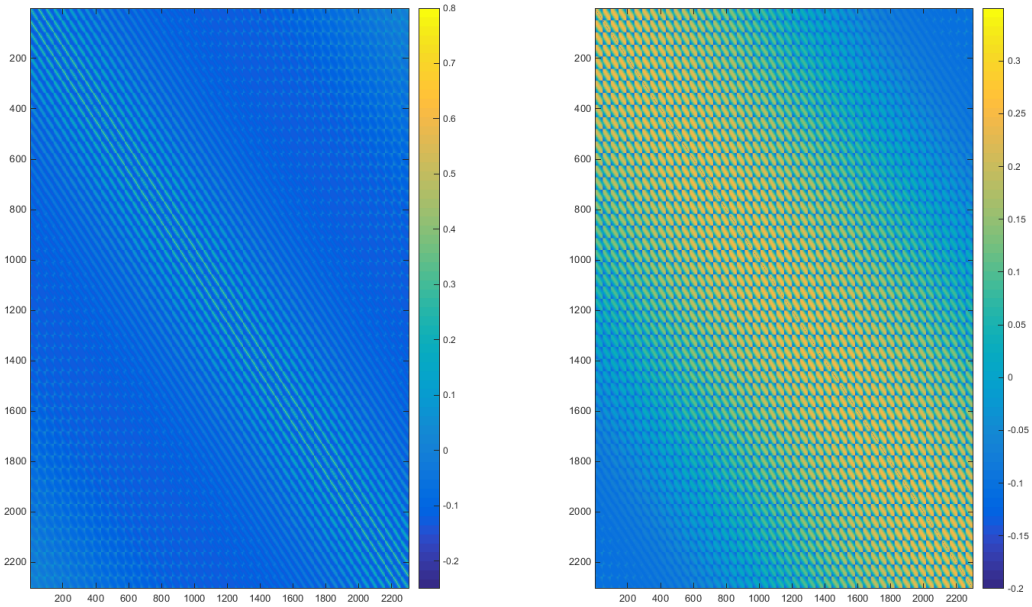
Recall from §2.2.3 that $\mathbb{S}_m^{-1} \approx \mathbb{G}^m$, i.e. the matrices of Type B are approximately equal to the matrices of Type A. For interest, by comparing Figures 4-10(b) and 4-10(a), we can see the entries of $\mathbb{S}_m^{-1}$ (Type B) and $\mathbb{G}^m$ (Type A) are qualitatively similar, apart from differences occurring at the edges within the PML region.

4.3.1.1 Wavespeeds

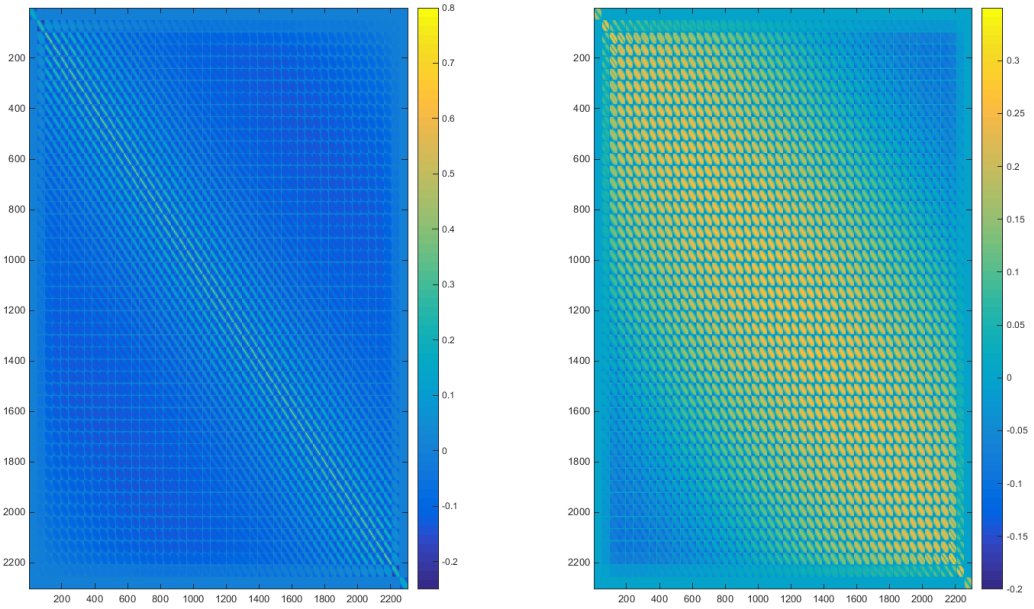
We consider a variety of wavespeed models in this thesis; some of the below wavespeeds are used in the subsequent numerics in §4.3.3-4.3.6, others not until §5, we state them together here for convenience. Again their shorthand notation is given in bold.

C1 Homogeneous $c(x) \equiv 1$.

CL Converging lens $c(x) = 1 - \exp(-32((x_1 - 1/2)^2 + (x_2 - 1/2)^2))$, see Figure



(a) A plot of entries of a Type A matrix when $N = 50$, real part in the left panel, imaginary part in the right panel.



(b) A plot of entries of a Type B matrix when $N = 50$, real part in the left panel, imaginary part in the right panel.

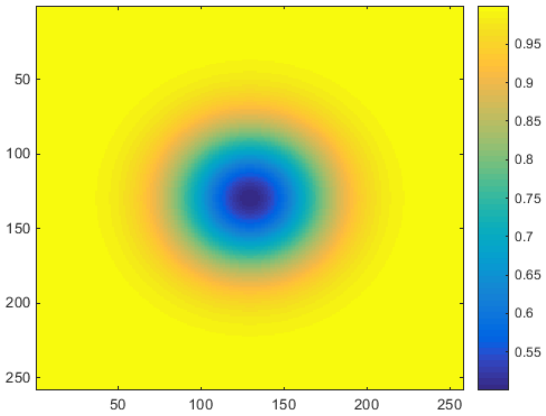


Figure 4-10: **CL** converging lens

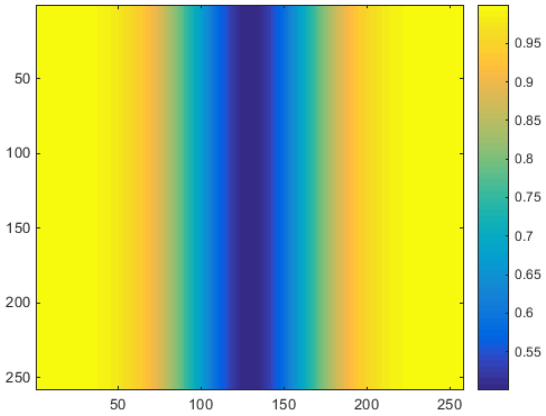


Figure 4-11: **CVW** vertical waveguide

4-10.

CVW Vertical waveguide $c(x) = 1 - 0.5\exp(-32((x_1 - 1/2)^2))$, see Figure 4-11.

CH* Half of domain one wavespeed, the other half another wavespeed, with contrast between the halves of *, for example for **CH3**, the contrast is 3. For the experiments in §4.3.3-4.3.6, c is 0.5 on $[0, 0.5] \times [0, 1]$ and 1.5 on $[0.5, 1] \times [0, 1]$, so that the contrast is 3.

CM For this wavespeed we use a 125x125 coarse sample of the Marmousi model, as illustrated in Figure 4-12. The Marmousi model is an artificial seismic data set created by the Institut Français du Pétrole [112]. It is based on a profile

of actual subsurface parameters from the Cuanza basin and is a widely used test data set for seismic imaging as it exhibits complicated features for solution methods, see for example the normal fault lines creating tilted regions in Figure 4-12. Even though we use only a coarse sample, this data still provides a higher contrast and more complicated wavespeed for our experiments.

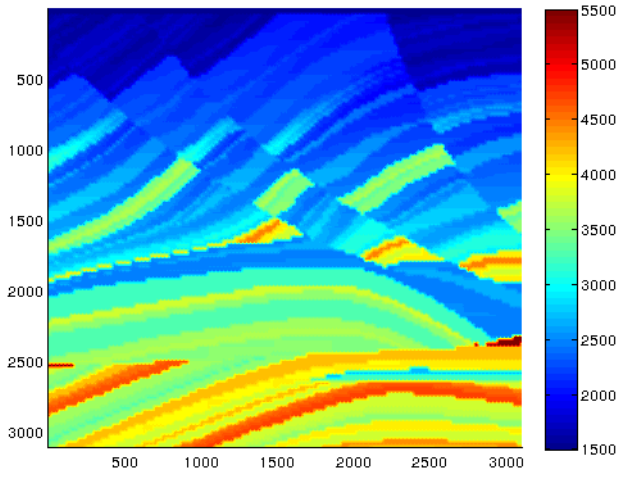


Figure 4-12: [101, Figure 5-12] Plot of $c(x)$ for the part of the Marmousi model used. The full Marmousi data set was created by Institut Français du Pétrole [112].

4.3.2 Recall ε-rank

To investigate the properties of low-rank approximations we being by constructing the SVD of matrix blocks of Types A-D, as this allows us to find a sufficient rank of approximation to get each entry to within ε. How this is achieved is explained through properties of the SVD seen in the next two theorems. (Note that low-rank approximations of matrix blocks would never be constructed in this way in practice as this is a very high cost method for finding them, $\mathcal{O}(mn^2)$ for an $m \times n$ matrix [50, §5.4.5 p239]. Instead one could use for example the randomized SVD algorithm in [66], however this is not necessary for our simple demonstrative test.)

Theorem 4.3.1. ***Singular Value Decomposition*** *Reproduced from [68, p580] Any matrix $A \in \mathbb{C}^{m' \times n'}$ has a singular value decomposition (SVD) $A = U'\Sigma V'^{*}$,*

$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbb{C}^{m' \times n'}$, $p = \min(m', n')$, *where* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$ *and* $U \in \mathbb{C}^{m' \times m'}$, $V \in \mathbb{C}^{n' \times n'}$ *are both unitary. The* σ_i *are the singular values of* A *and the columns of* U *and* V *are the left and right singular vectors of* A*, respectively.*

Theorem 4.3.2. *Given a matrix* $M \in \mathbb{C}^{m' \times n'}$ *and* $\varepsilon > 0$*, if* R *is such that*

$$\sum_{i=R+1}^{\min(m',n')} \sigma_i < \varepsilon, \tag{4.15}$$

then there exists $\mathbf{\Phi}_j \in \mathbb{C}^{m' \times 1}$ *and* $\mathbf{\Psi}_j \in \mathbb{C}^{n' \times 1}$ *for* $j \in \{1, \ldots, R\}$ *such that*

$$\left| \left(M - \sum_{j=1}^{R} \mathbf{\Phi}_j \mathbf{\Psi}_j^T \right)_{i,i'} \right| < \varepsilon, \tag{4.16}$$

for all $i \in \{1, \ldots, m'\}$ *and* $i' \in \{1, \ldots, n'\}$.

Proof. We define R as the minimum R to satisfy 4.15. Theorem 4.3.1 gives us a singular value decomposition for M where $u'_{i,j}$ and $v'_{i,j}$ are the i,jth entries of U' and V' respectively. Now we define the vectors

$$\begin{aligned} (\mathbf{\Phi}_j)_i &:= u'_{i,j}, && \text{for } j \in \{1, \ldots, \min\{m', n'\}\} \text{ and } i \in \{1, \ldots, m'\}, \\ (\mathbf{\Psi}_j)_i &:= \sigma_j v'_{i,j}, && \text{for } j \in \{1, \ldots, \min\{m', n'\}\} \text{ and } i \in \{1, \ldots, n'\}. \end{aligned}$$

If we find the first R pairs of these vectors satisfy 4.16 we are done. Note that $M = \sum_{j=1}^{\min(m',n')} \mathbf{\Phi}_j \mathbf{\Psi}_j^T$ by Theorem 4.3.1. Also since U' and V' are unitary matrices, their rows and columns form orthogonal bases for $\mathbb{C}^{m'}$ and $\mathbb{C}^{n'}$ respectively. This means any row or column has $\|\cdot\|_2 = 1$ and so each entry ≤ 1. Then

$$\begin{aligned} \left| M_{i,i'} - \sum_{j=1}^{R} \left(\mathbf{\Phi}_j \mathbf{\Psi}_j^T \right)_{i,i'} \right| &= \left| \sum_{j=R+1}^{\min(m',n')} \left(\mathbf{\Phi}_j \mathbf{\Psi}_j^T \right)_{i,i'} \right| \\ &\leq \sum_{j=R+1}^{\min\{m',n'\}} |\sigma_i| < \varepsilon \quad \text{by assumption.} \end{aligned}$$

□

The ranks of the low-rank approximations to off-diagonal matrix blocks reported in this section are the R as in (4.15). We shall call R the ε-rank.

4.3.3 Experiment 1

We investigate how the ε-rank changes as $\varepsilon \to 0$, when $k_I = 0$. From the expressions for the rank in (4.4) and (4.6), for admissible off-diagonal blocks of $\mathbb{G}^m$, we expect that the ε-rank to $\sim \log^2\left(\frac{1}{\varepsilon}\right)$ as $\varepsilon \to 0$ (when the factor containing ε dominates the maxima in (4.4) and (4.6)). The results for Types A-D are displayed in Figures 4-13, 4-14, 4-15 and 4-16 respectively.

We note the following.

- All four matrices have similar ε-ranks which suggests good translation of ε-rank properties from $\mathbb{G}^m$ to $\mathbb{S}_m^{-1}$. This is also born out by other experiments as their results are also very similar for matrix Types A-D, hence for future experiments we present only one plot.

- All four graphs show a good fit to the linear, least-squares best fit line with gradient of about 1. This suggests $\sim \log(1/\varepsilon)$ and our estimate of $\sim \log^2(1/\varepsilon)$ may actually be pessimistic.

- For Type A, the direct evaluation of the Green's function, the ε-rank fits the best fit line perfectly (and better than for the other types) and the best fit line's gradient is 1 to several significant figures. This is a consequence of the fact that the Types B and C are an approximation of Type A and this is a feature replicated amongst many of the future experiments (whose results we therefore do not present in full).

- Type C is very similar to both Types B and A with slightly lower ε-ranks than Type B.

- Type D is very similar to all the others, with slightly higher ε-ranks than Type B.

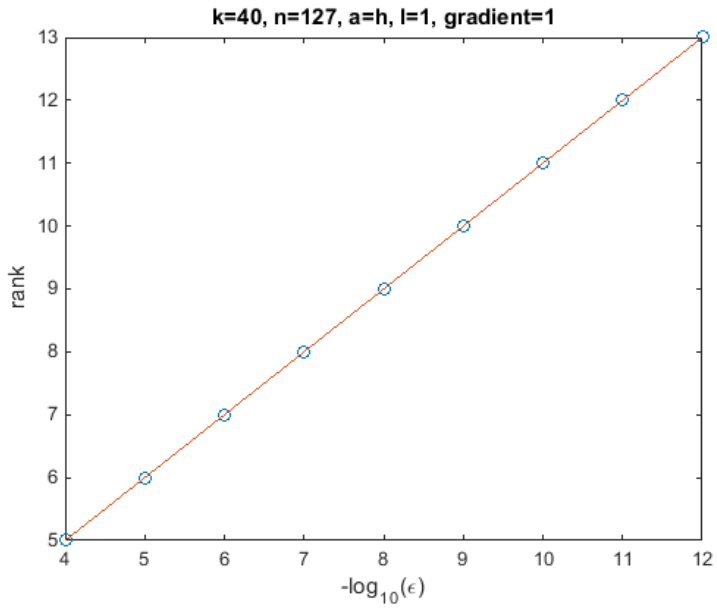


Figure 4-13: ε-rank of Type A matrix for decreasing ε. The line is a linear, least-squares best fit.

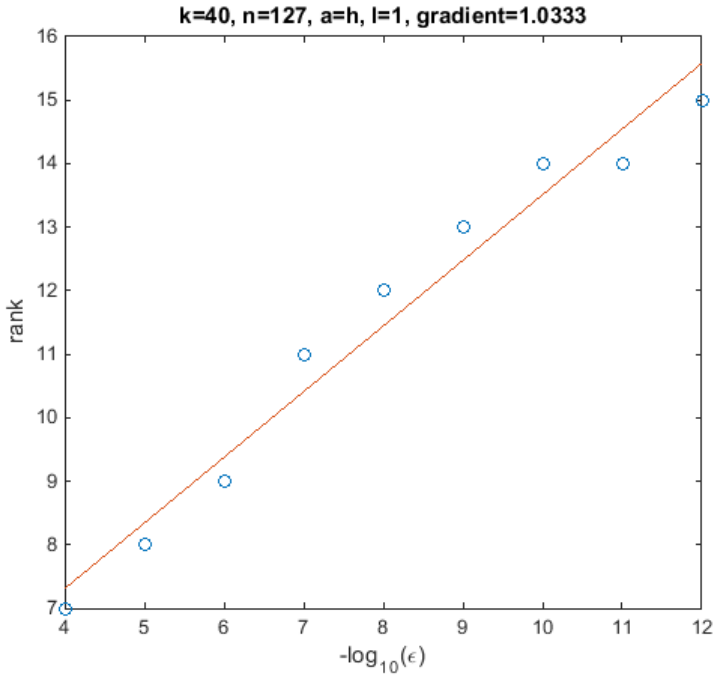


Figure 4-14: ε-rank of Type B matrix for decreasing ε. The line is a linear, least-squares best fit.

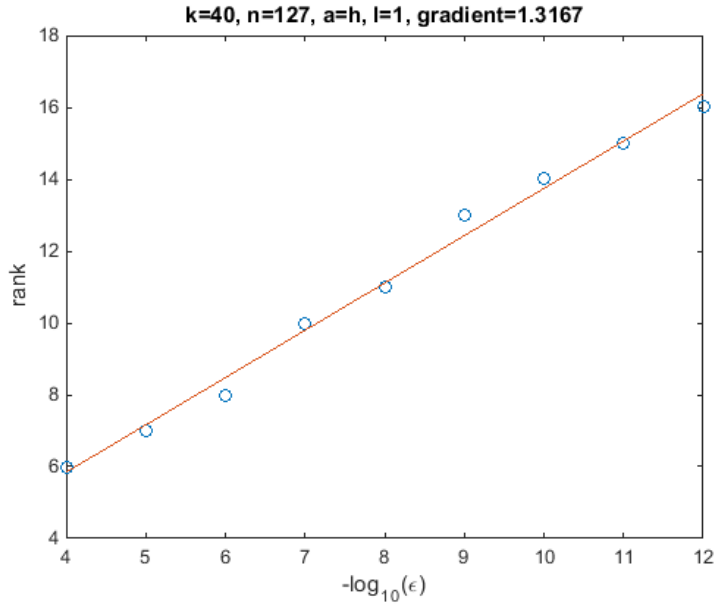


Figure 4-15: ε-rank of Type C matrix for decreasing ε. The line is a linear, least-squares best fit.

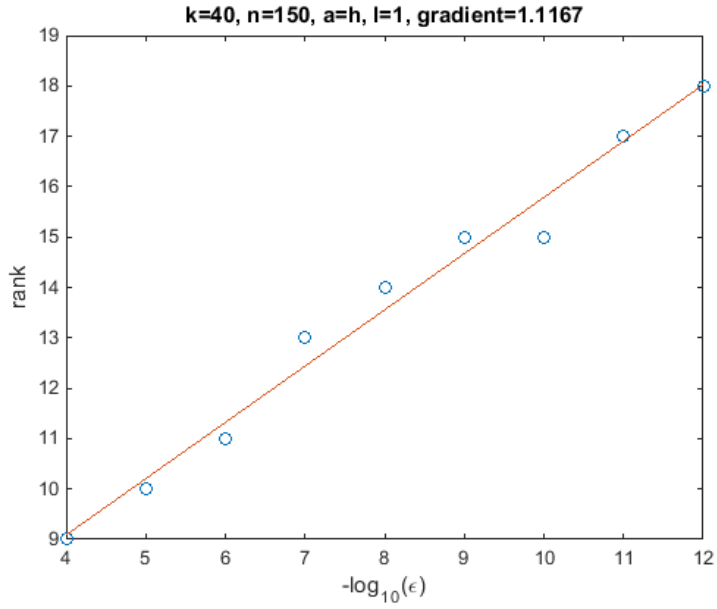


Figure 4-16: ε-rank of Type D matrix for decreasing ε. The line is a linear, least-squares best fit.

4.3.4 Experiment 2

We look at ε-ranks for different values of k_R and k_I. This allows us to verify both the k_R-independence we have predicted in the expressions for the rank in (4.4) and (4.8) and also that the ε-rank $\sim \exp(-2k_I a)$ for increasing k_I (when the factor containing ε dominates the maxima in (4.4) and (4.6)). The results are displayed in Table 4.3.4.

	$a = 0.2$						$a = h$					
	Type A			Type B			Type A			Type B		
k_R	20	30	40	20	30	40	20	30	40	20	30	40
k_I / n	81	129	174	81	129	174	81	129	174	81	129	174
0	5	5	5	7	7	7	10	11	12	12	13	14
$k_R^{0.25}$	5	5	5	6	6	6	10	11	11	11	12	13
$k_R^{0.5}$	4	4	4	5	4	4	10	11	11	10	11	11
$k_R^{0.75}$	4	3	2	3	3	2	9	9	9	8	9	9
$k_R^{0.9}$	3	2	0	3	2	0	8	8	8	8	8	8

Table 4.1: ε-rank of matrix blocks with varying k_R, k_I, n and separation a with $\varepsilon = 10^{-10}$. The results for Type C with $a = h$ are similar to Type B with slightly lower ε-ranks in some cases. The results for Type D with $a = h$ are similar to Type B (with $N \equiv 150$) with some ε-ranks higher and some lower. We did not perform experiments for Type C or D with $a = 0.2$.

We note the following.

- We see k-independence when $a = 0.2$ but not when $a = h$.
- In all cases the ε-rank decreases as absorption increases, but because the ε-ranks are so small and take discrete values it is not possible to verify if the rate of the decrease is exponential with k_I.

4.3.5 Experiment 3

In Experiment 3 we fix a, k_R and ε and use various different values of k_I to see if we can verify the improved quality of approximation predicted by (4.7). To investigate the improved quality of approximation we take the ε-rank when

$k_I = 0$, which we define to be R_0 and compute

$$s := \sum_{i=R_0+1}^{\text{end}} \sigma_i.$$

By Theorem 4.3.2, the low-rank approximation of rank R_0 approximates each entry of the matrix block to within s. Then we compute an improvement ratio ε/s, that we expect by (4.7) to $\sim \exp(-k_I a)$ for increasing k_I. The results are shown in Figures 4-17 and 4-18. (The results we have not presented for Type C are similar with slightly higher log(improvement ratio) values than Type B with slightly higher gradients of 1.0842 and 0.77678 respectively. The results we have not presented for Type D have gradients 1.0755 and 1.1234 respectively, with overall slightly lower improvement factors in $k = 20$ case and slightly higher in $k = 30$ case.)

We note the following.

- In all cases we have used $a = h$, which in theory should be too small a factor for improvements to be seen as absorption is added (since $k_I a \to 0$ as $k_R \to \infty$ when $a = h$ and $k_I < k_R$). However, improvements are clearly visible, so Theorem 4.2.33 is pessimistic in this regard. This means that the improvement in the quality of the approximation appears to be occurring in all far-field blocks of $\mathbb{G}^m$, even those close to the diagonal.

- The improvements show good fit to the linear least-squares trendlines so that the improvement ratio ε/s does $\sim \exp(k_I)$ as k_I increases, as we expected.

- The fact that the improvement ratios and gradients are greater in Types B and C relative to Type A show us that adding absorption benefits the FEM solutions more than the direct evaluation of the Green's function.

4.3.6 Experiment 4

Here we seek to verify Remark 3.3.5, i.e. that for certain δ, ν, μ combinations we get a low-rank separable expansion for fatter domains. This could provide some justification for 'sweeping' multiple rows of the grid at a time, to form

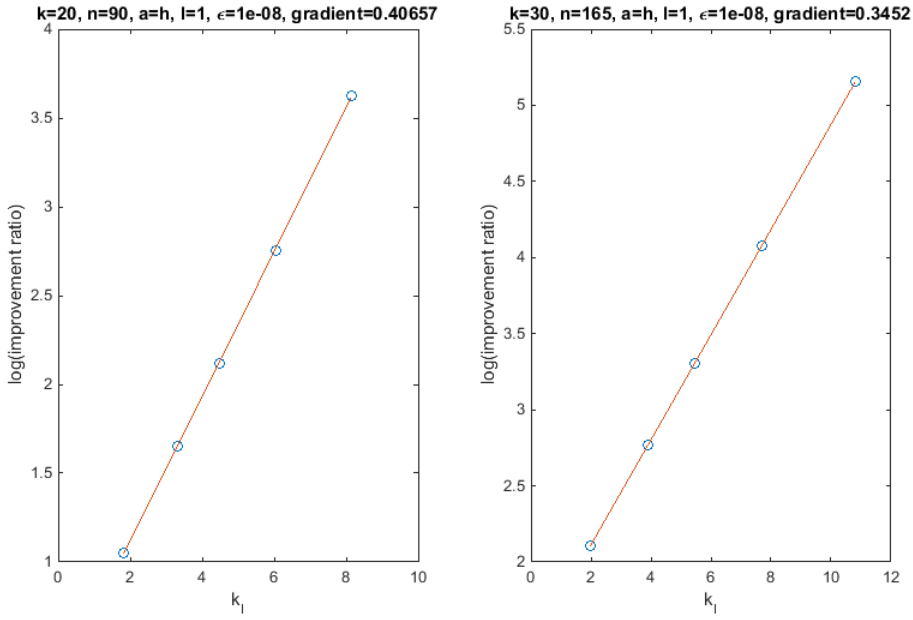


Figure 4-17: Improvement in quality of approximation of matrix blocks of Type A. Line a linear, least-squares best fit.

larger Schur complement matrices $\mathbb{S}_m^{-1}$ containing several rows that still admit good low-rank approximations. (This is despite the fact that, in §4.2, we did not develop a result for $\mathbb{G}^m$ based on Remark 3.3.5.)

First we do an experiment with $a = h$, which gives a set of δ, ν, μ values outside of our range. We see in Figure 4-19 the rank increases almost linearly with the number of rows, so it appears that our theory such as Remark 3.3.5 rightly considers this range of values to be a problem. (The result with Type B matrix blocks was similar.)

Next we do an experiment with $a = 0.4$ and $\delta = 0.75$, which means that ν is close to 0 and so even with the most pessimistic value of μ our $\frac{\delta-\nu/\mu}{2} > 0$ and $\nu < \delta/\mu$ requirements from Remark 3.3.5 are readily satisfied. The results in Figure 4-20 show the following.

- Although there is still some increase in the rank as the rows increase, it is far less pronounced, and so this shows partial verification of the need for Remark 3.3.5.

- Adding absorption in the right-hand panel further reduces the ranks seen in

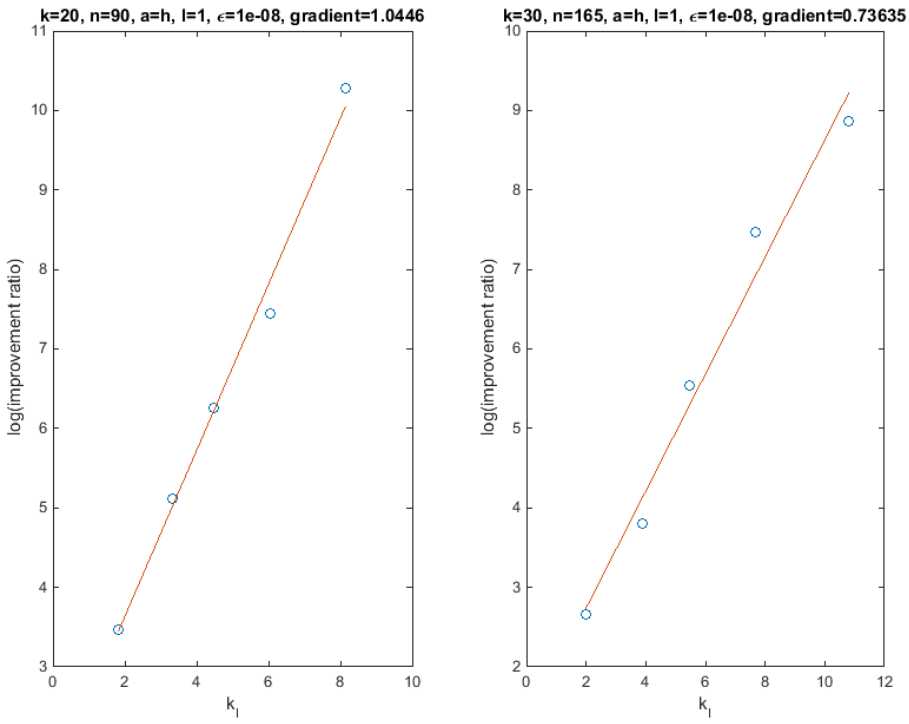


Figure 4-18: Improvement in quality of approximation for matrix blocks of Type B. Line a linear, least-squares best fit.

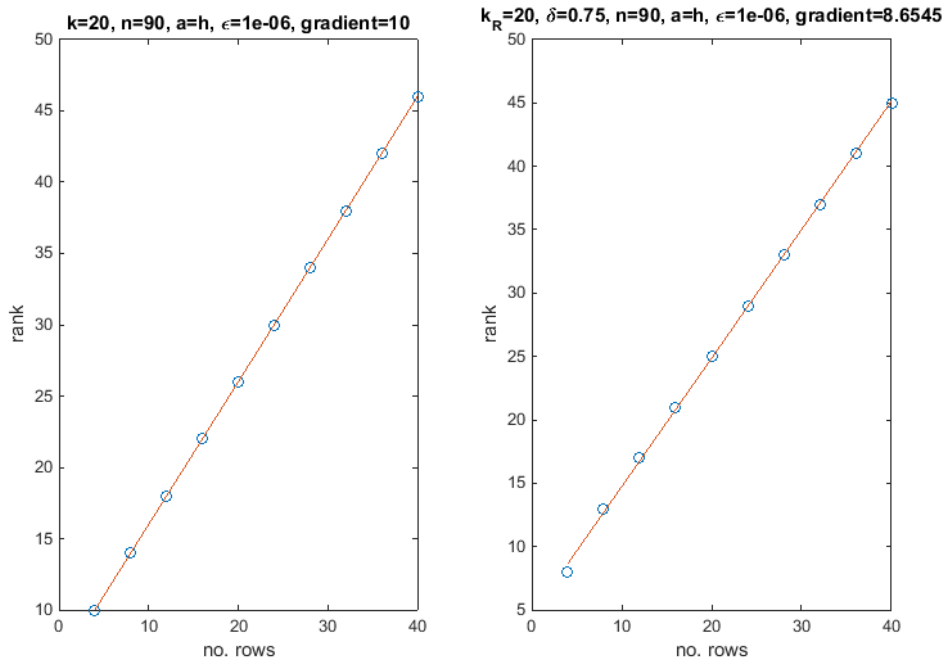


Figure 4-19: Improvement in quality of approximation of matrix blocks of Type A. Line a linear, least-squares best fit.

the left-hand panel. However, it is still the case that the dramatic reduction in the ranks is clearly visible without absorption, so that our theory such as Remark 3.3.5 may be pessimistic in requiring absorption to witness this phenomenon.

Note that the graph for Type B blocks was similar, with slightly higher ranks in the left-hand panel but occasionally slightly lower ranks in the right-hand panel, so that we again see slightly more benefits to adding absorption in the FEM matrix blocks than for the direct evaluation of the Green's function.

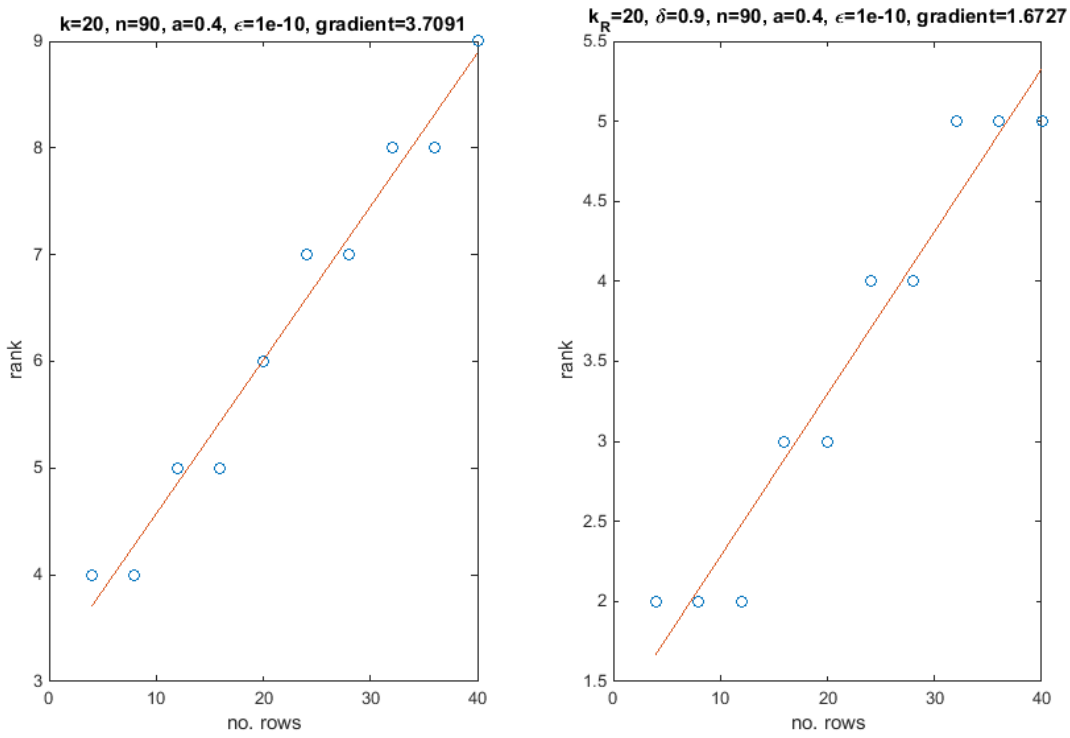


Figure 4-20: The ε-rank of matrix blocks of Type A containing increasing numbers of rows in the grid.

Chapter 5

Numerical Experiments on Sweeping Preconditioners with Absorption

5.1 Numerical Experiments

5.1.1 Description of Sweeping Preconditioner

We now do numerical experiments examining the effect of absorption in a sweeping preconditioner. We use Engquist and Ying's Sweeping Preconditioner [34, 35] for our experiments due to its close relation to the theory and experiments in §2-§4.

Engquist and Ying use two ways of approximating the Schur complements: using the moving PML method [35] and $\mathcal{H}$-matrices [34]. In §5.1.2-5.1.4 we outline these methods in detail as we present results of our numerical experiments using several variants of the sweeping preconditioner; in §5.1.2 we use the moving PML method and in §5.1.3-5.1.4 we use $\mathcal{H}$-matrices (recall we have seen $\mathcal{H}$-matrices in §1.8.3 and §1.8.3). In particular, in §5.1.3 we use weakly admissible $\mathcal{H}$-matrices (see §4.2.2, especially Definition 4.2.6) and in §5.1.4 we use strongly admissible $\mathcal{H}$-matrices (see §4.2.2, especially Definition 4.2.7).

In §5.1.2-5.1.4 we present results of numerical experiments using several variants of the sweeping preconditioner; the remainder of this section gives details about the experiments common to all of the variants (defining abbreviations as

shorthand notation for each of the experiments). The different wavespeed models we use were described in §4.3.1.1. Next we give details of the iterative solver, source terms and model parameters we use.

Table of numerical experiment abbreviations

Abbreviation	Type	Description	Section
C1	Wavespeed	Homogeneous	§4.3.1.1
CL	Wavespeed	Converging lens	§4.3.1.1
CVW	Wavespeed	Vertical waveguide	§4.3.1.1
FPS	Source	Gaussian point source	§5.1.1.2
FPW	Source	Plane wave	§5.1.1.2
HK1	n	$h \sim k_R^{-1}$	§5.1.1.3
HK1.5	n	$h \sim k_R^{-3/2}$	§5.1.1.3
PNA	Problem to be solved	$A\boldsymbol{u} = \boldsymbol{f}$	§5.1.1.4
PWA	Problem to be solved	$A_{\text{abs}}\boldsymbol{u}_{\text{abs}} = \boldsymbol{f}$	§5.1.1.4
ALT	Alternative Parameters	n, h, C and η changed	§5.1.1.6

Table 5.1: For reference we provide a list of all the abbreviations used to identify the numerical experiments. A description of the different preconditioners can be found in the following sections: the moving PML preconditioner in §5.1.2 and the weakly/strongly admissible preconditioners in §5.1.3 and §5.1.4 respectively. Details of the iterative solver and the PML parameters are given in §5.1.1.1 and §5.1.1.5-5.1.1.6 respectively.

5.1.1.1 Iterative Solver

We use the iterative solver GMRES [99,100] for all the numerical experiments in §5.1.1 (for further details about GMRES see §1.5.2). The GMRES starting guess for all experiments is the zero vector. The stopping criteria for all numerical experiments is when the relative residual of the ith iterate $\mathbf{u_i}$: $(\|\mathbf{f} - A\mathbf{u_i}\|/\|\mathbf{f}\|)$, is less than the tolerance 10^{-6}.

5.1.1.2 Source terms

We use two source terms.

FPS The Gaussian point source

$$f(x) = -\exp\left(-(1.5k_R - 1)^2\left[(x_1 - \acute{x}_1)^2 + (x_2 - \acute{x}_2)^2\right]\right),$$

with $\acute{x} = [\acute{x}_1, \acute{x}_2]^T = [1/2, 1/8]$.

FPW A plane wave multiplied by an exponentially decaying factor

$$f(x) = \exp(-8\pi k_R((x_1 - 1/8)^2 + (x_2 - 1/8)^2))\exp(\sqrt{2}\pi k_R i(x_1 + x_2)).$$

5.1.1.3 Discretisation Points n

We use the finite difference method described in §2.1.2. The number of degrees of freedom in the problem $N = n^2$, where n is the number of grid points per row.

We consider two different levels of grid refinement.

HK1 This level of refinement has $h \sim 1/k_R$, i.e. a constant number of points per wavelength. The number of points per wavelength is at least 8 and is a similar level of discretisation to [34, 35]. The values of n are

k_R	n
64	96
128	192
256	384
512	768

HK1.5 This level of refinement has $h \sim 1/k_R^{3/2}$. This is a finer level of discretisation than the constant points per wavelength above. This is in order to avoid the pollution effect, see §1.4.2. We make the discretisation finer at all levels, so when $k_R = 64$, the discretisation level is at least 10 points per wavelength.

k_R	n
64	128
128	384
256	1024

5.1.1.4 Problem

PNA The model problem with corresponding linear system $A\boldsymbol{u} = \boldsymbol{f}$, see Definition 1.9.2.

PWA A variant of the model problem where we include absorption in the wavenumber in the same way as in the preconditioner, i.e. $k = k_R + \mathrm{i}k_I$. This problem has corresponding linear system $A_{\mathrm{abs}}\boldsymbol{u}_{\mathrm{abs}} = \boldsymbol{f}$, and preconditioned system $\tilde{A}^{-1}_{\mathrm{abs}}A_{\mathrm{abs}}\boldsymbol{u} = \tilde{A}^{-1}_{\mathrm{abs}}\boldsymbol{f}$, (compare with Definition 1.9.3); note that $\boldsymbol{u}_{\mathrm{abs}}$ is less oscillatory than $\boldsymbol{u}$.

5.1.1.5 PML parameters

Recall the definition of the PMLs in §4.3.1. The values of the parameters for the PMLs are chosen for our numerical experiments as follows:

$$\eta = 12h,$$
$$C = 2/\mu,$$

so $\theta_1 = \theta_2 = \frac{1}{1+\mathrm{i}\phi(x)2\pi/k_R}$, where

$$\phi(x) = \begin{cases} \frac{1}{72h^2}\left(\frac{x-12h}{12h}\right)^2, & \text{if } 0 \le x \le 12h, \\ 0, & \text{if } 12h < x < 1-12h, \\ \frac{1}{72h^2}\left(\frac{x-1+12h}{12h}\right)^2, & \text{if } 1-12h \le x \le 1. \end{cases}$$

We use the same values in the artificial PMLs, so $D_{\mathrm{PML}} = 12$.

5.1.1.6 Alternative Set of Parameters

Recall the definition of the PMLs in §4.3.1. An alternative set of values of the parameters for the PMLs are chosen for our numerical experiments as follows:

$$\eta = 2/k,$$
$$C = 3\pi k,$$

so $\theta_1 = \theta_2 = \frac{1}{1+\mathrm{i}\phi(x)2\pi/k_R}$, where

$$\phi(x) = \begin{cases} \frac{3\pi k^2}{2}\left(\frac{x-2/k}{2/k}\right)^2, & \text{if } 0 \le x \le 2/k, \\ 0, & \text{if } 2/k < x < 1-2/k, \\ \frac{3\pi k^2}{2}\left(\frac{x-1+2/k}{2/k}\right)^2, & \text{if } 1-2/k \le x \le 1. \end{cases}$$

We do not use these parameters in the moving PML experiments.

HK1 This level of refinement has $h \sim 1/k_R$, i.e. a constant number of points per wavelength. The values of n are

k_R	n
16	128
32	256
64	512
128	1024

We do not consider finer grids with these alternative parameters.

5.1.2 Moving PML version

In this section we look at approximating the Schur complement matrices using the moving PML method as introduced in [35]. We include a few details to outline this method for completeness based on [35] and [101, §5]. We use code provided by Lexing Ying for our experiments. Aside from using different parameters (as given in 5.1.1.3-5.1.1.6) we make no changes to the method outlined in [35] or the code used that was provided by Lexing Ying.

Recall that in this section, we assume each Schur complement matrix corresponds to D rows of the grid, where D exactly divides n. (The algorithms we give can readily be adapted for Schur complements with differing numbers of corresponding rows D_m, as in Definition 2.2.2, but this makes the presentation more technical and we do not do it here.)

In [35] only the algorithm for Schur complements corresponding to single rows, with PMLs on 3 sides of the grid (corresponding to solving a half-plane Dirichlet problem, rather than a full-plane problem which requires PMLs on all 4 sides of the grid) is given in full: how to extend it to the multiple-line version and the version with PMLs on 4 sides of the grid are explained separately. Also

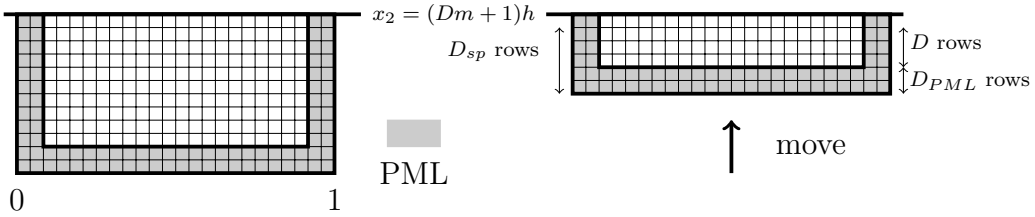


Figure 5-1: To approximate $\mathbb{S}_m^{-1}$ (corresponding to D rows), we use a method that moves the PML up and solves the half-plane problem on the right.

neither [35] nor [101, §5] give the costings for the multiple-line version. Therefore we give more details about the formulation of the algorithm for the version of the preconditioner with Schur complement matrices corresponding to multiple rows and with PMLs on 4 sides of the grid in §5.1.2.1. We also give the costings for this version in §5.1.2.1-5.1.2.2.

5.1.2.1 Approximating Schur complements

The idea of the moving PML method is to approximate multiplication by $\mathbb{S}_m^{-1}$ by efficiently solving a new Helmholtz problem on Ω_m (Definition 2.2.1 and Figure 2-6). We recall that $\mathbb{S}_m^{-1}$ is related to the half-plane problem in Figure 2-7 (as explained in §2.2.3) and that Ω_m is a small subdomain of the half-plane problem. We create a subdomain problem on each Ω_m by moving the PML of the half-plane problem up to the row $Dm - D$ as in Figure 5-1, hence the name 'moving PML method'. Then each subdomain problem is a half-plane problem with zero-Dirichlet condition on the upper boundary (the row of grid nodes above the top of Ω_m). The total depth of the subdomain problem is $D_{sp} = D + D_{PML}$, where D is the number of grid rows in $\mathbb{S}_m^{-1}$ and D_{PML} is the number of grid rows in the artificial PML, see Figure 5-1. As $D_{sp} \ll n$ we say the subdomain problem is quasi-1D.

Since we consider the case with PML on the top of the grid, the algorithm in fact performs a double sweeping motion: creating subproblems by sweeping down and moving the PML at the top down (as well as what we've already seen of sweeping up and moving the PML at the bottom up), see Figure 5-2.

To explain how the new quasi-1D problems (i.e. those in Figures 5-1 and 5-2) provide an approximation to $\mathbb{S}_m^{-1}$, we look at these more closely.

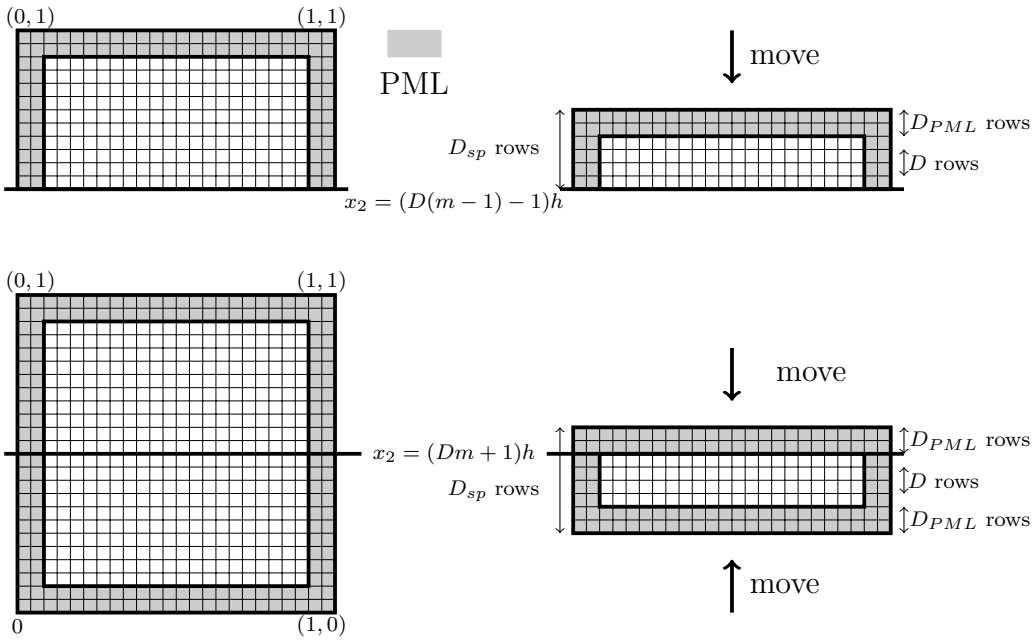


Figure 5-2: When there are PMLs on all sides of the grid the subdomain problems are created differently. Top: for the subdomains in the upper half of the grid we approximate $\mathbb{S}_m^{-1}$ (corresponding to D rows), by moving the PML *down* and solving the half-plane problem on the right. Bottom: for the middle subdomain we approximate $\mathbb{S}_m^{-1}$ (corresponding to D rows), by moving the top PML down *and* the bottom one up and solving the full-plane problem on the right.

Definition 5.1.1. ***(Quasi-1D problem matrix*** $\mathbb{H}_m$***)*** *Let* $\mathbb{H}_m$ *be the matrix arising from finite difference discretisation of the quasi-1D problems in Figures 5-1 and 5-2.*

Multiplication by $\mathbb{H}_m^{-1}$ gives an approximation to multiplying by $\mathbb{S}_m^{-1}$. To see the reasons for this, we first note that

$$\boldsymbol{v} = \mathbb{H}_m^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \boldsymbol{f}^m \end{pmatrix}, \tag{5.1}$$

where $\boldsymbol{v}$ is the solution to the quasi-1D problem.

To explain the reasoning behind the method of approximating the Schur complements in this version of the preconditioner we must first recall some details from §2.2.3. In Definition 2.2.18 we define $\widehat{\boldsymbol{u}}^m = A_m^{-1}\widehat{\boldsymbol{f}}^m$ to be the discretisation of a problem on the half-plane displayed in red in Figure 2-7. In Proposition 2.2.14 we see that the bottom right entry of A_m^{-1} is $\mathbb{S}_m^{-1}$. Recall $\widehat{\mathbf{u}}^m|_{\Omega_m}$ and $\widehat{\mathbf{f}}^m|_{\Omega_m}$ are $\widehat{\boldsymbol{u}}^m$ and $\widehat{\boldsymbol{f}}^m$ truncated to Ω_m respectively, then multiplying out the bottom right entry of A_m^{-1} gives

$$\widehat{\mathbf{u}}^m|_{\Omega_m} = \mathbb{S}_m^{-1}\widehat{\mathbf{f}}^m|_{\Omega_m}. \tag{5.2}$$

This shows that in some sense multiplication by $\mathbb{S}_m^{-1}$ corresponds to partially solving a Helmholtz half-plane problem in Definition 2.2.18, taking a source concentrated on Ω_m and then observing the solution on Ω_m.

Another key component of the moving PML preconditioners is to do multiplication by $\mathbb{H}_m^{-1}$ efficiently. Note that $\mathbb{H}_m^{-1}$ can be transformed into a D_{sp} banded matrix by reordering the nodes in $\hat{\Omega}_m$; so that they are ordered in the x_2 direction as in Figure 5-3. Let P_m be the permutation matrix induced by this new node ordering. Then $P_m H_m P_m^T$ is a banded matrix with D_{sp} upper and lower diagonals. Due to the banded structure, the decomposition

$$L_m U_m = P_m \mathbb{H}_m P_m^T, \tag{5.3}$$

can be calculated efficiently and then multiplication by $\mathbb{H}_m^{-1}$ can be calculated efficiently by Gaussian elimination.

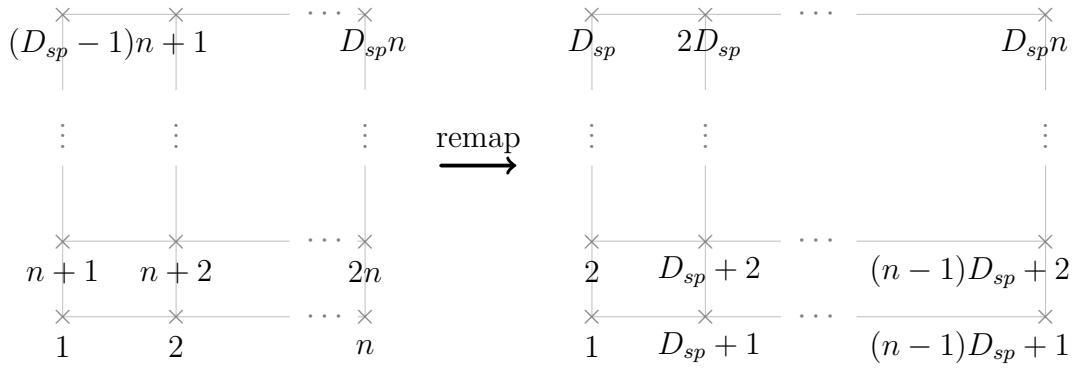


Figure 5-3: The $n \times D_{sp}$ set of nodes ordered lexicographically in the x_1 direction (left) and x_2 direction (right).

In the next two lemmas we find the cost of finding P_m, L_m and U_m from (5.3) and the cost of solving systems with matrix $\mathbb{H}_m$ (i.e. multiplying by $(P_m^T L_m U_m P_m)^{-1}$) , which are novel compared to [35, 101] in the sense that they are explicit in terms of D, D_{sp} and D_{PML}.

Lemma 5.1.2. *Calculating P_M, L_m, U_m for $\mathbb{H}_m$ for $m \in \{1, \ldots, M\}$ takes $\mathcal{O}\left(\frac{D_{sp}^4}{D} N\right)$ operations.*

Proof. Finding the permutation matrices P_m and $L_m U_m$ decompositions of $n \times n$, $\bar{D}$-banded matrices is $\mathcal{O}(\bar{D}^3 n)$ (this can be derived from first principles or looking at the costings for the equivalent algorithm where $D = 1$ on [35, p694]). Our $\mathbb{H}_m$ matrices are size $D_{sp}n \times D_{sp}n$ and D_{sp} banded, so it costs $\mathcal{O}(D_{sp}^4 n)$ to calculate the permutation matrices P_m and $L_m U_m$ decomposition for each $\mathbb{H}_m$. There are n/D $\mathbb{H}_m$ matrices, giving the overall cost $\mathcal{O}\left(\frac{D_{sp}^4}{D} n^2\right) = \mathcal{O}\left(\frac{D_{sp}^4}{D} N\right)$. □

Lemma 5.1.3. *Multiplying by $P_m^T U_m^{-1} L_m^{-1} P_m$ for all $m \in \{1, \ldots, M\}$ by Gaussian elimination can be done in $\mathcal{O}\left(\frac{D_{sp}^3}{D} N\right)$ operations.*

Proof. For size $n \times n$, $\bar{D}$ banded matrices, multiplying by $P_m^T U_m^{-1} L_m^{-1} P_m$ for any m is $\mathcal{O}(\bar{D}^2 n)$ (by looking at [35, Algorithm 2.4, p694], the equivalent algorithm when $D = 1$). Therefore for our size $D_{sp}n \times D_{sp}n$, D_{sp} banded matrices the cost of multiplying by $P_m^T U_m^{-1} L_m^{-1} P_m$ for any m is $\mathcal{O}(D_{sp}^3 n)$. As $M = \frac{n}{D}$, the overall cost is $\mathcal{O}\left(\frac{D_{sp}^3}{D} N\right)$. □

We note that the following lemma for the memory costs only considers the cost of storing P_m, L_m and U_m, as this implicitly defines the inverse $P_m^T U_m^{-1} L_m^{-1} P_m$

needed in the preconditioner (recall from §1.6 that only the action of multiplying by the preconditioning matrix is needed, the matrix itself is not needed and indeed this is how Engquist and Ying's algorithm is formulated, see [35, Algorithm 2.3 and 2.4]).

Lemma 5.1.4. *An upper bound on the memory cost of storing P_m, L_m, U_m for $\mathbb{H}_m$ for $m \in \{1, \ldots, M\}$ is $\mathcal{O}\left(\frac{D_{sp}}{D} N\right)$.*

Proof. The $\mathbb{H}_m$ matrices (and therefore P_M, L_m and U_m matrices) are size $D_{sp}n \times D_{sp}n$. The $\mathbb{H}_m$ matrices are D_{sp} banded, so they have $\mathcal{O}(D_{sp}n)$ non-zero entries. The L and U matrices in the LU decomposition of a D_{sp} banded matrix inherit the same D_{sp} banding, see for example [50, Theorem 4.3.1]. The P_m permutation matrices contain exactly n non-zero entries. Therefore the memory costs to store the permutation matrices P_m and the $L_m U_m$ decomposition for each $\mathbb{H}_m$ is $\mathcal{O}(D_{sp}n)$. There are n/D $\mathbb{H}_m$ matrices, giving the overall cost as $\mathcal{O}\left(\frac{D_{sp}}{D} n^2\right) = \mathcal{O}\left(\frac{D_{sp}}{D} N\right)$. □

5.1.2.2 Algorithm

The moving PML algorithm we use in our experiments is the same as that used in the experiments in [35] and [101]. Recall that the algorithm is essentially performing the matrix-multiplication (2.14), where multiplication of the Schur complements is performed efficiently, as outlined in §5.1.2.1 (i.e. the $\mathbb{S}_m^{-1}$ matrices are approximated with LU decompositions of the permuted quasi-1D problems matrices $\mathbb{H}_m$ and the order of the multiplication of the $\mathbb{S}_m^{-1}$ matrices in (2.14) is changed to a double sweeping motion). For full details of the algorithm we refer the reader to [35, §2.3]. In contrast to the experiments in [35] and [101] our experiments focus on investigating the effect of absorption on the preconditioner, so we use differing levels of absorption and also consider various values of D.

By Lemmas 5.1.2 and 5.1.3 the overall computational cost of solving our linear system using this approximation to A^{-1} as a preconditioner is $\mathcal{O}(N_I^2 D_{sp}^4 N)$ where N_I is the number of GMRES iterations. In most of our experiments with optimal parameters and those conducted in [35, 101], N_I is observed to depend at most logarithmically on N, i.e., $N_I \lesssim \mathcal{O}(\log(N))$. Therefore, the moving PML preconditioner provides a way to solve our model problem in roughly $\mathcal{O}(\log^2(N)\ D_{sp}^4 N)$

operations, close to the ideal $\mathcal{O}(N)$ operations. An upper bound for the memory costs is obtained by summing the memory costs for the GMRES algorithm (see (1.29)) and for the storage of the preconditioning matrix in Lemma 5.1.4 to obtain $\mathcal{O}\left(\frac{D_{sp}}{D}N + N_I N + N_I^2\right) = \mathcal{O}\left(\frac{D_{sp}}{D}N_I N\right) = \mathcal{O}\left(\frac{D_{sp}}{D}\log(N)N\right)$, using the empirical observation that $N_I \lesssim \mathcal{O}(\log(N))$. Again, this is close to the ideal $\mathcal{O}(N)$ memory costs.

5.1.2.3 Numerical Results

We conduct numerical experiments with the moving PML method according to the description given earlier in §5.1.1. Solutions corresponding to some of the experiments are given in Figures 5-4(a) - 5-4(f).

Next we give iteration counts for all of the experiments. For all the following experiments we look at the two levels of grid refinement **HK1** and **HK1.5** (even and odd numbered tables respectively). First we conduct an experiment with homogeneous wavespeed, point source and absorption in the problem (**C1**, **FPS** and **PWA**) in Tables 5.2 - 5.3. Then we conduct experiments with homogeneous wavespeed, no absorption in the problem and with point source and plane wave solutions in turn (**C1**, **PNA** and **FPS**/**FPW** respectively) in Tables 5.4 - 5.7. Finally we conduct experiments with no absorption in the problem, point source and with varying wavespeed models, the converging lens and the vertical waveguide (**PNA**, **FPS** and **CL**/**CVW** respectively) in Tables 5.8-5.11.

5.1.2.4 Interpretation of moving PML preconditioner results

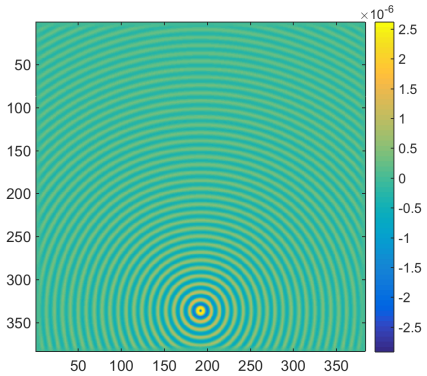
We discuss several aspects of this moving PML results in turn.

1) Connection to previous low-rank results

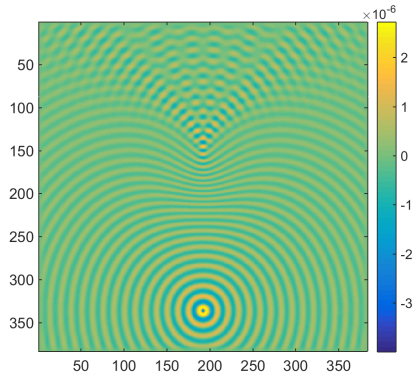
2) Reduction in iteration counts as D increases

3) Little change in iteration counts when finer grids are used

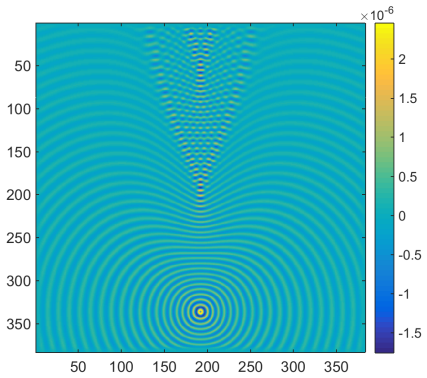
4) Improvements due to absorption



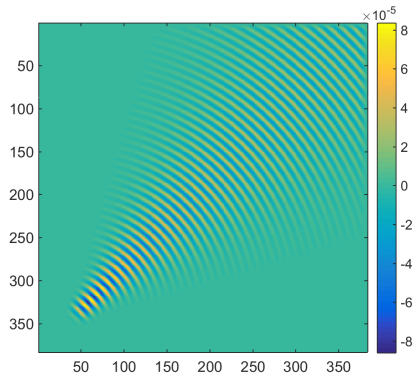
(a) Moving PML solution u, $k_R = 256$. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**



(b) Moving PML solution u, $k_R = 256$. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CL**.



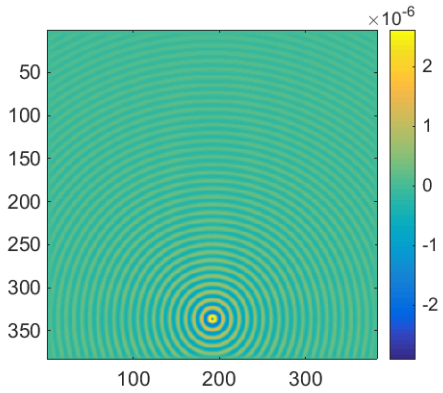
(c) Moving PML solution u, $k_R = 256$. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CVW**.



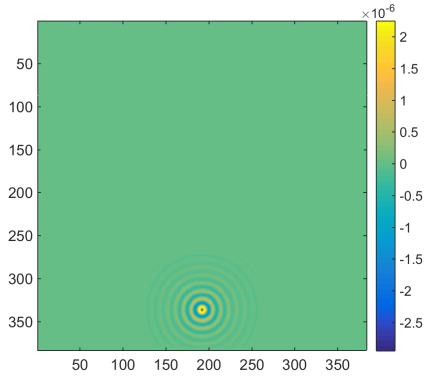
(d) Moving PML solution u, $k_R = 256$. Dependence of h on k_R: **HK1**. Source: **FPW**. Absorption level: **PNA**. Wavespeed model: **C1**.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	5	5	5	5	5	4	4	4	4	4	3
128	6	6	6	6	6	5	5	5	4	4	3
256	9	9	8	8	8	8	8	7	4	4	3
512	18	17	17	16	15	14	14	11	5	4	3
$D = 6$											
64	4	4	4	4	4	4	4	4	3	3	3
128	5	5	5	5	5	5	5	4	4	3	3
256	7	7	7	7	7	7	7	6	4	3	3
512	14	13	13	12	12	12	11	9	5	3	3
$D = 12$											
64	4	4	4	4	4	4	4	4	3	3	3
128	5	5	5	5	5	5	5	4	4	3	3
256	7	6	6	6	6	6	6	6	4	3	3
512	10	9	9	9	9	9	8	7	4	3	3
$D = 18$											
64	4	4	4	4	4	4	4	4	3	3	3
128	4	4	4	4	4	4	4	4	3	3	3
256	6	6	6	6	6	6	6	6	4	3	3
512	8	8	8	8	8	7	7	7	4	3	3
$D = 24$											
64	4	4	4	4	3	3	3	3	3	3	3
128	4	4	4	4	4	4	4	4	3	3	3
256	6	6	5	5	5	5	5	5	4	3	3
512	8	8	7	7	7	7	7	6	4	3	3

Table 5.2: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS** Absorption level: **PWA**. Wavespeed model: **C1**.



(e) Moving PML solution u, $k_R = 256$. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PWA**. Wavespeed model: **C1**. $k_I = 1$.



(f) Moving PML solution u, $k_R = 256$. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PWA**. Wavespeed model: **C1**. $k_I = k_R^0.5$.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	5	5	4	4	4	4	4	4	4	3	3
128	6	6	6	6	6	5	5	5	4	3	3
256	9	9	8	8	8	8	8	7	4	3	3
$D = 6$											
64	4	4	4	4	4	4	4	4	3	3	3
128	6	5	5	5	5	5	5	5	4	3	3
256	8	8	8	8	8	8	7	7	4	3	3
$D = 12$											
64	4	4	4	4	4	4	4	4	3	3	3
128	5	5	5	5	5	5	5	4	4	3	3
256	8	8	7	7	7	7	7	6	4	3	3
$D = 18$											
64	4	4	4	4	3	3	3	3	3	3	3
128	5	5	4	4	4	4	4	4	3	3	3
256	7	7	7	6	6	6	6	5	4	3	3
$D = 24$											
64	4	4	4	4	4	4	4	4	3	3	3
128	5	5	5	5	5	5	5	4	4	3	3
256	7	6	6	6	6	6	6	6	4	3	3

Table 5.3: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS** Absorption level: **PWA**. Wavespeed model: **C1**.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	5	5	6	7	7	8	9	10	17	34	*
128	6	6	7	7	7	8	9	11	22	*	*
256	9	8	8	9	9	9	10	12	27	*	*
512	18	17	16	16	15	14	14	15	35	*	*
$D = 6$											
64	4	5	6	7	7	8	9	10	17	34	*
128	5	6	6	7	7	8	9	11	22	*	*
256	7	7	8	8	8	9	10	12	28	*	*
512	14	13	13	12	12	12	12	14	35	*	*
$D = 12$											
64	4	5	6	7	7	8	9	10	17	34	*
128	5	6	6	7	7	8	9	11	22	*	*
256	7	7	7	7	8	9	9	12	28	*	*
512	10	9	9	10	10	10	10	14	35	*	*
$D = 18$											
64	4	5	6	7	7	8	9	10	17	34	*
128	4	5	6	7	7	8	9	11	22	*	*
256	6	6	7	7	8	9	9	12	27	*	*
512	8	8	8	8	8	9	10	14	35	*	*
$D = 24$											
64	4	5	6	7	7	8	9	10	17	34	*
128	4	5	6	7	7	8	9	11	22	*	*
256	6	6	6	7	7	8	9	12	27	*	*
512	8	8	8	8	8	9	10	14	35	*	*

Table 5.4: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS** Absorption level: **PNA**. Wavespeed model: **C1** * indicates did not converge within 50 iterations.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	5	5	6	7	7	8	9	10	17	34	*
128	6	6	7	7	7	8	9	11	21	*	*
256	9	9	9	9	9	9	10	12	27	*	*
$D = 6$											
64	4	5	6	7	7	8	9	10	17	34	*
128	6	6	6	7	7	8	9	11	21	*	*
256	8	8	8	9	9	9	10	12	27	*	*
$D = 12$											
64	4	5	6	7	7	8	9	10	17	34	*
128	5	6	6	7	7	8	9	11	21	*	*
256	8	8	8	8	9	9	10	12	27	*	*
$D = 18$											
64	4	5	6	7	7	8	9	10	17	34	*
128	5	6	6	7	7	8	9	11	21	*	*
256	7	7	7	8	8	9	9	12	27	*	*
$D = 24$											
64	4	5	6	7	7	8	9	10	17	34	*
128	5	6	6	7	7	8	9	11	21	*	*
256	7	7	7	7	8	9	9	12	27	*	*

Table 5.5: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS** Absorption level: **PNA**. Wavespeed model: **C1** * indicates did not converge within 50 iterations.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	5	6	6	7	8	9	9	11	19	38	*
128	6	6	7	7	8	9	10	12	23	*	*
256	9	9	9	9	9	10	11	13	30	*	*
512	17	16	15	15	15	14	14	16	39	*	*
$D = 6$											
64	4	6	6	7	8	9	9	11	19	38	*
128	5	6	7	7	8	9	10	12	23	*	*
256	7	7	7	8	8	9	10	13	31	*	*
512	12	11	11	11	11	11	11	15	40	*	*
$D = 12$											
64	4	6	6	7	8	9	9	11	19	38	*
128	5	6	7	7	8	9	10	12	23	*	*
256	6	6	7	7	8	9	10	13	30	*	*
512	8	8	8	8	9	9	10	14	40	*	*
$D = 18$											
64	4	6	6	7	8	9	9	11	19	38	*
128	4	6	6	7	8	9	10	12	23	*	*
256	5	6	7	7	8	9	10	13	30	*	*
512	7	7	7	8	8	9	10	14	40	*	*
$D = 24$											
64	4	6	6	7	8	9	9	11	19	38	*
128	4	6	6	7	8	9	10	12	23	*	*
256	5	6	7	7	8	9	10	13	30	*	*
512	6	7	7	8	8	9	10	14	40	*	*

Table 5.6: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPW** Absorption level: **PNA**. Wavespeed model: **C1** * indicates did not converge within 50 iterations.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	5	6	6	7	8	8	9	11	18	38	*
128	6	6	7	7	8	9	10	12	23	*	*
256	9	9	9	10	10	10	11	14	30	*	*
$D = 6$											
64	4	6	6	7	8	8	9	11	18	38	*
128	6	6	7	7	8	9	10	12	23	*	*
256	9	9	9	9	9	10	11	13	30	*	*
$D = 12$											
64	4	6	6	7	8	8	9	11	18	38	*
128	5	6	7	7	8	9	10	12	23	*	*
256	8	8	8	8	9	10	10	13	30	*	*
$D = 18$											
64	4	6	6	7	8	8	9	11	18	38	*
128	5	6	7	7	8	9	10	12	23	*	*
256	7	7	7	8	8	9	10	13	29	*	*
$D = 24$											
64	4	6	6	7	8	8	9	11	18	38	*
128	5	6	6	7	8	9	10	12	23	*	*
256	6	6	7	7	8	9	10	13	30	*	*

Table 5.7: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPW** Absorption level: **PNA**. Wavespeed model: **C1** * indicates did not converge within 50 iterations.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	6	6	7	7	8	9	9	11	18	38	*
128	8	8	8	8	8	9	10	12	23	*	*
256	12	12	12	12	11	11	12	13	29	*	*
512	32	30	28	26	25	23	22	19	38	*	*
$D = 6$											
64	5	6	7	7	8	9	10	11	18	38	*
128	7	7	7	8	8	9	10	12	23	*	*
256	10	10	10	10	10	10	11	13	30	*	*
512	22	21	20	19	19	17	16	16	38	*	*
$D = 12$											
64	5	6	7	7	8	9	10	11	18	38	*
128	6	7	7	7	8	9	10	12	23	*	*
256	8	8	9	9	9	9	10	13	30	*	*
512	14	13	13	13	12	12	12	15	38	*	*
$D = 18$											
64	5	6	6	7	8	9	10	11	18	38	*
128	6	6	7	7	8	9	10	12	23	*	*
256	8	8	8	8	9	9	10	13	30	*	*
512	11	10	10	10	11	11	11	14	38	*	*
$D = 24$											
64	4	6	6	7	7	9	9	11	18	38	*
128	6	6	7	7	8	9	10	12	23	*	*
256	7	7	8	8	8	9	10	13	30	*	*
512	10	10	10	10	10	10	11	14	38	*	*

Table 5.8: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS** Absorption level: **PNA**. Wavespeed model: **CL** * indicates did not converge within 50 iterations.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	6	6	7	7	8	9	9	11	18	37	*
128	8	8	8	8	8	9	10	12	23	*	*
256	12	12	12	12	12	12	12	14	29	*	*
$D = 6$											
64	6	6	7	7	8	9	10	11	18	37	*
128	8	8	8	8	8	9	10	12	23	*	*
256	12	12	11	11	11	11	11	13	29	*	*
$D = 12$											
64	6	6	7	7	8	9	9	11	18	37	*
128	7	7	7	8	8	9	10	12	23	*	*
256	11	11	10	10	11	11	11	13	29	*	*
$D = 18$											
64	5	6	6	7	7	9	9	11	18	37	*
128	6	6	7	7	8	9	10	12	23	*	*
256	9	9	9	9	9	9	10	13	29	*	*
$D = 24$											
64	5	6	6	7	8	9	9	11	18	37	*
128	7	7	7	7	8	9	10	12	23	*	*
256	9	9	9	9	9	9	10	13	29	*	*

Table 5.9: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS** Absorption level: **PNA**. Wavespeed model: **CL** * indicates did not converge within 50 iterations.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	6	7	7	8	8	10	11	12	22	46	*
128	8	8	8	8	9	10	11	14	29	*	*
256	11	11	11	11	11	11	12	16	38	*	*
512	24	23	22	21	20	19	19	20	50	*	*
$D = 6$											
64	6	7	7	8	9	10	11	12	22	46	*
128	8	8	8	9	9	10	11	14	29	*	*
256	10	10	10	10	10	11	12	16	38	*	*
512	17	16	16	15	15	15	15	18	50	*	*
$D = 12$											
64	6	7	7	8	9	10	11	12	22	46	*
128	8	8	8	9	9	10	11	14	29	*	*
256	10	10	10	10	11	11	12	16	38	*	*
512	14	14	14	14	14	14	15	18	50	*	*
$D = 18$											
64	6	7	7	8	9	10	11	12	22	46	*
128	8	8	8	8	9	10	11	14	29	*	*
256	10	10	10	10	10	11	12	16	38	*	*
512	14	14	14	14	14	14	14	18	50	*	*
$D = 24$											
64	5	6	7	7	8	10	11	12	22	46	*
128	8	8	8	8	9	10	11	14	29	*	*
256	10	10	10	10	10	11	12	16	38	*	*
512	13	13	13	13	13	14	14	18	50	*	*

Table 5.10: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS** Absorption level: **PNA**. Wavespeed model: **CVW** * indicates did not converge within 50 iterations.

k_R & D \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$D = 1$											
64	6	7	7	8	8	9	10	12	21	44	*
128	7	7	8	8	9	10	10	13	26	*	*
256	11	10	10	10	10	11	11	14	34	*	*
$D = 6$											
64	6	7	7	8	8	9	10	12	21	44	*
128	7	7	8	8	9	10	10	13	26	*	*
256	10	10	10	10	10	11	11	14	34	*	*
$D = 12$											
64	6	6	7	8	8	9	10	12	21	44	*
128	7	7	8	8	9	10	11	13	26	*	*
256	9	9	9	10	10	10	11	14	34	*	*
$D = 18$											
64	6	6	7	7	8	9	10	12	21	44	*
128	7	7	8	8	9	10	11	13	26	*	*
256	9	9	9	9	9	10	11	14	34	*	*
$D = 24$											
64	6	6	7	8	8	9	10	12	21	44	*
128	7	7	8	8	9	10	11	13	26	*	*
256	9	9	9	10	10	10	11	14	34	*	*

Table 5.11: Moving PML preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CVW** * indicates did not converge within 50 iterations.

5.1.2.5 Interpretation of Moving PML Results

1) Connection to previous low-rank results

We are especially interested in interpreting the result of the moving PML preconditioner experiments in terms of the low-rank results in §3 - 4. The quasi-1D problems solved as part of the moving PML method are mostly half-plane Helmholtz problems for which we have low-rank results (§3-4).

However, the connection between the moving PML preconditioner and the low-rank results is less clear than it is for the other preconditioners that we consider in §5.1.3-5.1.4. This is due to the way the Schur complements are approximated. The other preconditioners we experiment with create low-rank approximations of off diagonal blocks of the Schur complements that clearly correspond to the low-rank results on separated domains from §4. In contrast, the moving PML preconditioner approximates the quasi-1D problems exactly.

The difference in the form of the approximation of the moving PML method from the low-rank results is an important though subtle distinction. To illustrate this, consider that a similar looking change of making the subdomains taller in each has an 'opposite' effect. The quality of the low-rank approximation tends to decrease with increasing d (see §3), but the quality of the preconditioner to increase with increasing D. The former we know from §3, as d increases, the increased information in the Hankel function on the larger domains must still be contained in the low-rank separable expansion. For the reasons the quality of the approximation increases with increasing D, see **2)** below.

2) Reduction in iteration counts as D increases

A feature common to all the numerical experiments (see Tables 4.1-4.12) is the decrease in the iteration counts with increasing D (at least for lower values of k_I where the difference between the problems being solved and being used to form the preconditioner is not the dominant factor, see **P1)** in §1.9.2.2).

Recall D is the number of grid rows in each Ω_m and therefore the number of grid rows in each quasi-1D problem (not including the artificial PML, see Figure 5-1). Each of the quasi-1D problems is solved exactly by Gaussian elimination, the only approximation then being that we are solving the small quasi-1D problems as approximations to the half-plane problems of Figure 2-7. Therefore the larger D is, the less of an approximation the preconditioner is making and the better the preconditioner will perform. In fact, if $D = n$ the subproblem becomes

the original problem (ignoring the top PML) and so we exactly calculate A^{-1}, although at an impractically high cost.

3) Little change in iteration counts when finer grids are used

Recall that we consider two levels of grid discretisation, **HK1** and **HK1.5**, as we expect that the solutions on the coarser grids are affected by the pollution effect, described in §1.4.2. However, as can be seen by comparing any odd and even numbered pair of tables, the iteration counts do not change significantly between the two levels of grid refinement. Where there is a difference, the iteration counts for the finer grid are a little higher.

The reasons for this occasional and small difference are not clear. Possibly the coarse grids are so coarse that the solution is less accurate due to the pollution effect, in which case the overall behaviour and therefore difficulty of solving the problem may be altered slightly.

4) Improvements due to absorption

For the experiments with absorption in the problem **PWA** we see the expected decrease in iterations as k_I increases in all cases, see Tables 5.2 and 5.3. As demonstrated for $\mathcal{H}$-matrix approximations of Schur complements in §4, the more absorption the better the approximation is, recall **P2)** from §1.9. The factor of improvement we theoretically expect to be $\exp(k_I)$, as seen in for example §3.2.3.1 and Theorem 4.2.33. However, it is not so easy to verify this explicitly from the experiments.

For the experiments without absorption in the problem (but still in the preconditioner) **PNA** we still see some improvements due to absorption in the iteration counts. The improvements are only seen for $k_I \in \{0.25, \ldots, 2\}$ (apart from Table 5.8 for various cases of $k_R = 512$). We expect the improvements to be visible only for smaller values of k_I, due to the interplay we saw between the requirements **P1)** and **P2)** from §1.9 *"to balance the conflicting needs of **P1)** and **P2)** on the [amount of absorption], we expect that some, but not too much absorption..., may be of benefit in reducing the number of GMRES iterations"*.

The improvements due to absorption for **PNA** are also only seen for $k_R = 256+$ and tend to appear more often for lower values of D. These facts suggest that benefits due to absorption are more likely to show through in the more difficult cases, either when the problem is harder (e.g. for higher wavenumbers) or when the preconditioner is a less good approximation to the inverse (as is the

case when D is smaller, see **2)**). Presumably in the other cases the approximation is already so good the absorption has little effect in improving the approximation and may therefore even be solely detrimental (as is observed sometimes) as the problems being solved and underlying the preconditioner diverge (see **P1)** and **P2)** from §1.9 again). We look more at this phenomenon in §5.1.3.2 **4)**.

5.1.3 Weakly Admissible $\mathcal{H}$-Matrix version

In this section we perform numerical experiments with Engquist and Ying's sweeping preconditioner, with weakly admissible $\mathcal{H}$-matrix approximation. We use code provided by Lexing Ying. Aside from using different parameters (as given in §5.1.1.3-5.1.1.6) we make no changes to the method outlined in [34]. The second variant of Engquist and Ying's sweeping preconditioner uses $\mathcal{H}$-matrices to approximate the Schur complement matrices. As we explained in §1.8.3, the preconditioner is created by approximating multiplication by A^{-1} using the decomposition (2.14) in the Hierarchical Matrix Framework (HMF).

We now give the algorithms of the preconditioner from [34]. The algorithms feature HMF operations hinv, hdiagmul and hmatvec. (We do not explain the exact variants of the HMF operations used in our numerical experiments here, as that is of little added value to this thesis and we refer the reader to [34]). We state the algorithms for the cases $D \geq 1$, which is novel in the sense that it includes $D > 1$. A different version of the $\mathcal{H}$-matrix operations is needed when $D > 1$, due to the different structure of $\mathcal{H}$-matrices that are based on 2D geometry.

Algorithm 5.1.5. *[34, Algorithm 2.5, notation adapted]* ***Construction of $\mathcal{H}$-matrix approximations to Schur complement matrices $\mathbb{S}_m^{-1}$.***

$\mathcal{H}$-matrix approximations to the matrices $\mathbb{S}_m^{-1}$, or $\mathcal{H}$-matrices which store exactly the diagonal/tridiagonal matrices $\mathbb{A}_{m-1,m}$, $\mathbb{A}_{m,m}$, $\mathbb{A}_{m,m-1}$ are denoted by $\tilde{\mathbb{S}}_m^{-1}$, $\tilde{\mathbb{A}}_{m-1,m}$, $\tilde{\mathbb{A}}_{m,m}$ and $\tilde{\mathbb{A}}_{m,m-1}$ respectively.

1: $\tilde{\mathbb{S}}_1 = \mathbb{S}_1$. $\tilde{\mathbb{S}}_1^{-1} = \text{hinv}(\tilde{\mathbb{S}}_1)$.

2: for $m = 2, \ldots, M$

3: $\tilde{\mathbb{A}}_{m,m} = \mathbb{A}_{m,m}$, $\tilde{\mathbb{A}}_{m-1,m} = \mathbb{A}_{m-1,m}$ *and* $\tilde{\mathbb{A}}_{m,m-1} = \mathbb{A}_{m,m-1}$.

4: $\tilde{\mathbb{S}}_m = \text{hsub}(\tilde{\mathbb{A}}_{m,m}, \text{hdiagmul}(\tilde{\mathbb{A}}_{m,m-1}, \text{hdiagmul}(\tilde{\mathbb{S}}_{m-1}^{-1}, \tilde{\mathbb{A}}_{m-1,m})))$.

5: $\tilde{\mathbb{S}}_m^{-1} = \text{hinv}(\tilde{\mathbb{S}}_m)$.

6: end for

Engquist and Ying give the cost of Algorithm 5.1.5 when $D = 1$ as

$$\mathcal{O}(R^2 n^2 \log^2 n) = \mathcal{O}(R^2 N \log^2 N).$$

(Recall that R is the rank of the approximations of all the admissible off-diagonal blocks in the $\mathcal{H}$-matrix.) The preconditioner is now the approximation to multiplying by A^{-1} which uses the $\mathcal{H}$-matrix approximations to the $\mathbb{S}_m^{-1}$ matrices created in Algorithm 5.1.5. The preconditioner is given by the following algorithm that assumes the approximations to the Schur complements have already been created.

Algorithm 5.1.6. *[34, Algorithm 2.6, notation adapted]* ***Computation of*** $\boldsymbol{u} \approx A^{-1}\boldsymbol{f}$ ***using the approximate [Schur complement matrices] in the Hierarchical Matrix Framework.***

1: for $m = 1, \ldots, M$

2: $\boldsymbol{u}^m = \boldsymbol{f}^m$.

3: end for

4: for $m = 1, \ldots, M-1$

5: $\boldsymbol{u}^{m+1} = \boldsymbol{u}^{m+1} - \mathbb{A}_{m+1,m} \times \text{hmatvec}(\tilde{\mathbb{S}}_m^{-1}, \boldsymbol{u}^m)$.

6: end for

7: for $m = 1, \ldots, M$

8: $\boldsymbol{u}^m = \text{hmatvec}(\tilde{\mathbb{S}}_m^{-1}, \boldsymbol{u}^m)$.

9: end for

10: for $m = M-1, \ldots, 1$

11: $\boldsymbol{u}^m = \boldsymbol{u}^m - \text{hmatvec}(\tilde{\mathbb{S}}_m^{-1}, \mathbb{A}_{m,m+1}\boldsymbol{u}^{m+1})$.

12: end for

Engquist and Ying give the cost of Algorithm 5.1.6 when $D = 1$ as

$$\mathcal{O}(Rn^2 \log n) = \mathcal{O}(RN \log N).$$

The rank R has to be chosen with some care, but Theorem 2.2.23 suggests that as k increases R should only need to grow weakly to get the same accuracy of approximation. The total cost of solving the linear system is then $\mathcal{O}(N_I RN \log N)$ where N_I is the number of GMRES iterations (for full details see [34]). The numerical results in [34, §3] *"demonstrate that N_I is in practice very small, thus resulting in... [a solution through computation of] almost linear complexity"* as desired.

To approximate multiplication by (2.14) in the HMF, we should have reason to suppose $\mathcal{H}$-matrix approximations to the $\mathbb{S}_m^{-1}$ matrices and the $\mathbb{A}_{m+1,m}$ matrices (see (2.14) and (2.15)) will be good approximations.

$\mathcal{H}$-matrix approximations of $\mathbb{A}_{m+1,m}$ are exact as they are diagonal matrices, so that all their entries are within the diagonal blocks that are stored densely in $\mathcal{H}$-matrices.

Relation to previous theory and experiments approximating off-diagonal blocks of Schur complements in §4

We have looked extensively at effectively approximating the Schur complements using $\mathcal{H}$-matrices in §4. However, it is worth noting that the previous exploratory theory and experiments considered approximation of the Schur complements directly. (Recall that we considered direct approximation of the Schur complements by $\mathcal{H}$-matrices theoretically using their connection to $\mathbb{G}^m$ (see §4.2.4). We considered direct numerical approximation of the Schur complements from the exact Schur complements using an SVD (see §4.3).) Therefore, we have not previously considered the effect of 'recursively approximating' the Schur complements via their definition in Definition 2.2.5, as is done in Algorithm 5.1.5. However, the theory and experiments still motivate this recursive approximation.

5.1.3.1 Numerical Results

We conduct numerical experiments with the weakly admissible HMF preconditioner according to the description given earlier in §5.1.3. As we are solving the same problems as those in the moving PML preconditioner, the solutions are the same as those already seen in Figures 5-4(a) - 5-4(f) (apart from the case with the alternative PML parameters **ALT**).

Next we give iteration counts for all of the experiments. For each problem we consider different values of the approximation rank R (instead of sweeping different numbers of rows D as we did for the moving PML preconditioner).

Firstly, we choose a leaf-size of 12 and perform all the same combinations of experiments as for the moving PML method. So, as we did previously, for each problem we look at the two levels of grid refinement **HK1** and **HK1.5** (odd and even numbered tables respectively). We conduct an experiment with homogeneous wavespeed, point source and absorption in the problem as well as in the preconditioner (**C1**, **FPS** and **PWA**) in Tables 5.12 - 5.13. Then we conduct experiments with homogeneous wavespeed, no absorption in the problem and with point source and plane wave solutions in turn (**C1**, **PNA** and **FPS**/**FPW** respectively) in Tables 5.14 - 5.17. Finally we conduct experiments with no absorption in the problem, point source and with varying wavespeed models: the converging lens and the vertical waveguide (**PNA**, **FPS** and **CL**/**CVW** respectively) in Tables 5.18-5.21.

Secondly, we conduct some experiments with different leaf-sizes. (We do not always look at the two levels of grid refinement **HK1** and **HK1.5** for these experiments.) For our base problem with no absorption in the problem, homogeneous wavespeed and point source (**HK1**, **C1**, **FPS** and **PNA**) we consider the leaf-sizes **24** and **48** in Tables 5.22 and 5.23 respectively. We also do our base problem three times, with a different alteration each time: first for the grid refinement **HK1.5** in Table 5.24, then the plane wave **FPW** in Table 5.25 and then the vertical waveguide **CVW** in Table 5.26.

Finally, we conduct some experiments with the alternative set of PML parameters **ALT** described in §5.1.1.6 in Tables 5.27- 5.30. In these four tables, we respectively do the base problem (**HK1**, **C1**, **FPS** and **PNA**) for the leaf-sizes **24** and **48**, then change the wavespeed to **CL**/**CVW** for leaf-size **48**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$R = 3$											
64	4	4	4	4	4	4	4	4	3	3	3
128	5	5	5	5	5	5	4	4	4	3	3
256	6	6	6	6	5	5	5	5	4	3	3
512	8	7	7	7	7	7	6	6	4	3	3
$R = 5$											
64	3	3	3	3	3	3	3	3	3	2	2
128	3	3	3	3	3	3	3	3	3	2	2
256	4	4	4	4	4	4	4	3	3	2	2
512	4	4	4	4	4	4	4	4	3	3	2
$R = 10$											
64	2	2	2	2	2	2	2	2	2	2	2
128	3	3	3	3	3	3	3	3	2	2	2
256	3	3	3	3	3	3	3	3	2	2	2
512	3	3	3	3	3	3	3	3	3	2	2

Table 5.12: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PWA**. Wavespeed model: **C1**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$R = 3$											
64	5	4	4	4	4	4	4	4	3	3	3
128	6	6	6	6	6	6	6	5	4	3	3
256	11	10	10	10	10	9	9	8	5	4	3
$R = 5$											
64	3	3	3	3	3	3	3	3	3	2	2
128	4	4	4	4	4	4	4	3	3	3	2
256	4	4	4	4	4	4	4	4	3	3	3
$R = 10$											
64	2	2	2	2	2	2	2	2	2	2	2
128	3	3	3	3	3	3	3	3	2	2	2
256	3	3	3	3	3	3	3	3	3	2	2

Table 5.13: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS**. Absorption level: **PWA**. Wavespeed model: **C1**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	6	6	7	8	9	9	11
128	5	6	7	7	8	9	10	12
256	6	6	7	8	8	9	10	14
512	8	7	8	8	9	10	11	15
$R = 5$								
64	3	5	6	7	7	8	9	10
128	3	5	6	7	7	8	9	11
256	4	5	6	7	7	8	9	12
512	4	5	6	7	7	9	9	14
$R = 10$								
64	2	5	6	7	7	8	9	10
128	3	5	6	7	7	8	9	11
256	3	5	6	7	7	8	9	12
512	3	5	6	7	7	8	9	14

Table 5.14: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	5	6	6	7	7	8	9	10
128	6	7	7	8	8	9	10	12
256	11	11	11	11	11	12	13	15
$R = 5$								
64	3	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	4	6	6	7	7	8	9	12
$R = 10$								
64	2	5	6	7	7	8	9	10
128	3	5	6	7	7	8	9	11
256	3	5	6	7	7	8	9	12

Table 5.15: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS.** Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	6	7	7	8	9	10	11
128	5	6	7	8	8	9	10	13
256	6	6	7	8	8	10	11	14
512	7	7	7	8	9	10	11	16
$R = 5$								
64	3	6	6	7	8	9	9	11
128	4	6	6	7	8	9	9	12
256	4	6	6	7	8	9	10	13
512	4	6	7	7	8	9	10	14
$R = 10$								
64	2	6	6	7	8	9	9	11
128	3	6	6	7	8	9	9	12
256	3	6	6	7	8	9	10	13
512	3	6	7	7	8	9	10	14

Table 5.16: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPW**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	5	6	7	7	8	9	10	11
128	6	7	7	8	8	10	10	13
256	10	10	10	10	11	12	12	16
$R = 5$								
64	3	6	6	7	8	8	9	11
128	4	6	6	7	8	9	9	12
256	4	6	7	7	8	9	10	13
$R = 10$								
64	2	6	6	7	8	8	9	11
128	3	6	6	7	8	9	9	12
256	3	6	6	7	8	9	9	13

Table 5.17: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPW** Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	5	6	6	7	8	9	10	11
128	6	6	7	7	8	9	10	12
256	7	7	7	8	8	9	10	13
512	8	8	8	8	9	10	11	15
$R = 5$								
64	4	5	6	7	7	9	9	11
128	4	5	6	7	8	9	10	12
256	5	6	7	7	8	9	10	13
512	6	6	7	7	8	9	10	14
$R = 10$								
64	3	5	6	7	8	9	10	11
128	3	5	6	7	8	9	10	12
256	3	5	6	7	8	9	10	13
512	4	6	6	7	8	9	10	14

Table 5.18: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CL**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	5	6	7	7	8	9	9	11
128	7	7	8	8	9	9	10	12
256	10	10	11	11	11	12	13	16
$R = 5$								
64	4	5	6	7	7	9	9	11
128	4	5	6	7	8	9	9	12
256	6	6	7	7	8	9	10	13
$R = 10$								
64	3	5	6	7	7	9	9	11
128	3	5	6	7	8	9	10	12
256	4	5	6	7	8	9	10	13

Table 5.19: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS** Absorption level: **PNA**. Wavespeed model: **CL**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	6	7	8	8	10	11	12
128	5	6	7	8	9	10	11	14
256	6	6	7	8	9	10	11	16
512	7	7	8	9	9	10	12	18
$R = 5$								
64	3	6	7	7	8	10	11	12
128	4	6	7	8	8	10	11	14
256	4	6	7	8	9	10	11	16
512	4	6	7	8	9	10	11	17
$R = 10$								
64	3	6	7	7	8	10	11	12
128	3	6	7	8	8	10	11	14
256	3	6	7	8	9	10	11	16
512	4	6	7	8	9	10	11	17

Table 5.20: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CVW**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	5	6	7	8	8	10	11	12
128	6	7	7	8	9	10	11	14
256	11	11	11	11	12	12	13	17
$R = 5$								
64	3	6	7	7	8	9	10	12
128	4	6	7	7	8	9	10	13
256	4	6	7	7	8	9	10	14
$R = 10$								
64	3	6	7	7	8	9	10	12
128	3	6	7	7	8	9	10	13
256	3	6	7	7	8	9	10	14

Table 5.21: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS** Absorption level: **PNA**. Wavespeed model: **CVW**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	6	6	7	8	9	9	11
128	5	6	7	7	8	9	10	12
256	6	6	7	8	8	9	10	14
512	8	8	8	8	9	10	11	15
$R = 4$								
64	3	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	4	6	6	7	8	9	10	13
512	5	6	7	7	8	9	10	14
$R = 5$								
64	3	5	6	7	7	8	9	10
128	3	5	6	7	7	8	9	11
256	4	5	6	7	7	8	9	12
512	4	5	6	7	7	9	9	14
$R = 10$								
64	2	5	6	7	7	8	9	10
128	3	5	6	7	7	8	9	11
256	3	5	6	7	7	8	9	12
512	3	5	6	7	7	8	9	14

Table 5.22: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **24**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	5	6	7	7	8	9	10
128	5	6	7	7	8	9	10	12
256	6	6	7	8	8	9	10	14
512	7	7	8	8	9	10	11	15
$R = 4$								
64	3	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	4	6	6	7	8	9	10	13
512	5	6	7	7	8	9	10	14
$R = 5$								
64	3	5	6	7	7	8	9	10
128	3	5	6	7	7	8	9	11
256	4	5	6	7	7	8	9	12
512	4	5	6	7	7	9	9	14

Table 5.23: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **48**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	5	6	7	7	8	9	10
128	6	6	7	8	8	9	10	12
256	11	11	11	11	11	12	13	16
$R = 4$								
64	3	5	6	7	7	8	9	10
128	4	6	6	7	7	8	9	11
256	6	6	7	7	8	9	10	13
$R = 5$								
64	3	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	4	6	6	7	7	8	9	12

Table 5.24: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **48**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	6	6	7	8	9	10	11
128	5	6	7	7	8	9	10	12
256	6	6	7	8	8	10	11	14
512	7	7	7	8	9	10	11	16
$R = 4$								
64	3	6	6	7	8	9	9	11
128	4	6	6	7	8	9	10	12
256	4	6	7	7	8	9	10	13
512	5	6	7	7	8	9	10	15
$R = 5$								
64	3	6	6	7	8	9	9	11
128	3	6	6	7	8	9	9	12
256	4	6	6	7	8	9	10	13
512	4	6	7	7	8	9	10	14

Table 5.25: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPW**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **48**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
64	4	6	7	8	8	9	11	12
128	5	6	7	8	9	10	11	14
256	6	6	7	8	9	10	11	16
512	7	7	8	9	9	10	12	18
$R = 4$								
64	3	6	7	7	8	10	11	12
128	4	6	7	8	8	10	11	14
256	4	6	7	8	9	10	11	16
512	5	6	7	8	9	10	11	17
$R = 5$								
64	3	6	7	7	8	10	11	12
128	4	6	7	8	8	10	11	14
256	4	6	7	8	9	10	11	16
512	4	6	7	8	9	10	11	17

Table 5.26: Weakly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CVW**. Size smallest block: **48**.

$k_R/2\pi$ & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
16	8	8	9	9	10	12	14	14
32	12	12	12	12	12	14	15	16
64	18	17	16	16	17	17	18	20
128	29	28	27	26	26	26	26	28
$R = 5$								
16	5	6	8	9	10	12	13	13
32	8	8	8	9	10	12	14	15
64	11	11	11	12	12	13	15	17
128	20	19	19	18	18	18	19	22
$R = 10$								
16	4	6	8	9	10	12	13	13
32	5	6	8	9	10	12	14	15
64	7	7	8	9	10	12	14	17
128	12	12	12	12	13	14	15	19

Table 5.27: Weakly admissible preconditioner iteration counts. **ALT**. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **12**.

$k_R/2\pi$ & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
16	5	6	8	9	10	12	13	13
32	8	8	9	10	11	12	14	15
64	13	12	12	13	13	14	16	18
128	30	23	21	21	21	21	22	24
$R = 5$								
16	5	6	8	9	10	12	13	13
32	7	8	8	9	10	12	14	15
64	12	11	11	12	12	13	15	17
128	20	19	19	19	19	19	19	22
$R = 10$								
16	4	6	8	9	10	12	13	13
32	5	6	8	9	10	12	14	15
64	7	7	8	9	10	12	14	17
128	12	12	12	12	13	14	15	19

Table 5.28: Weakly admissible preconditioner iteration counts. **ALT**. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **48**.

$k_R/2\pi$ & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
16	6	6	7	8	9	11	12	12
32	8	8	8	9	10	11	12	13
64	12	12	12	12	12	13	14	16
128	19	18	18	19	19	19	19	21
$R = 5$								
16	5	6	7	8	9	11	12	12
32	7	7	8	8	9	11	12	13
64	9	9	9	10	10	11	13	15
128	13	12	13	13	13	14	14	17
$R = 10$								
16	4	6	7	8	9	11	12	12
32	5	6	7	8	9	11	12	13
64	7	7	7	8	9	11	12	15
128	9	9	9	9	10	11	13	16

Table 5.29: Weakly admissible preconditioner iteration counts. **ALT**. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CL**. Size smallest block: **48**.

$k_R/2\pi$ & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 3$								
16	8	8	10	12	14	17	20	20
32	9	10	11	13	15	18	22	25
64	13	13	14	16	17	20	24	30
128	23	22	22	22	22	25	28	37
$R = 5$								
16	6	8	10	12	14	17	20	20
32	8	8	11	13	14	18	22	25
64	11	11	13	14	16	20	24	30
128	16	16	16	17	18	22	25	36
$R = 10$								
16	4	8	10	12	14	17	20	20
32	6	8	10	12	14	18	22	25
64	7	9	11	13	15	19	23	30
128	10	11	12	14	16	20	24	35

Table 5.30: Weakly admissible preconditioner iteration counts. **ALT**. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **CVW**. Size smallest block: **48**.

5.1.3.2 Interpretation of weakly admissible preconditioner results

We discuss several aspects of the weakly admissible preconditioner results in turn.

1) Reduction in iteration counts as R increases

2) Little change in iteration counts for increased leaf-size

3) Higher iteration counts in the **ALT** cases

4) Small improvements due to absorption occasionally, more common in **ALT** case

We also note that we again observe that there is little change in the iteration counts when finer grids are used, apart from occasional small increases for the finer grids and that this was commented upon in 3) in §5.1.2.4.

1) Reduction in iteration counts as R increases

We see the behaviour of the preconditioner with respect to the changing rank R is similar for all the various experiments. Recall that the rank R denotes the rank of the approximations of each admissible block. We expect the preconditioner to perform better for larger ranks R.

Generally the numerical experiments do indeed show reduced iteration counts for larger R. We do not include results for $R = 1$ and $R = 2$ as the iteration counts were often large and unstable or the iterative method did not converge at all.

For $R = 4$ and 5 the preconditioner becomes much more stable and the iteration counts are all below 10 (when $k_I \leq 1$) in the non-**ALT** cases. Hence a rank of 4 or 5 is sufficient to stably generate a good approximate inverse. A rank of 4 or 5 would be good to choose in practice: we see that increasing the rank to 10 results in little change to the iteration counts for the increased memory and storage costs.

2) Little change in iteration counts for increased leaf-size

We compare iteration counts for different leaf-sizes. The experiment **HK1**, **FPS**, **PNA**, **C1** was conducted with three leaf-sizes: **12**, **24** and **48**, in Tables 5.14, 5.22 and 5.23 respectively. The iteration counts are almost identical for all the ranks in the tables. The main exception is the slight difference of the ranks

when $R = 3$, $k_R = 512$, $k_I = 0, 0.25$, however $R = 3$ is not in the stable range of R for this variant of the preconditioner (see **1)**).

3) Higher iteration counts in the ALT cases

In the cases with the alternative PML parameters, the iteration counts are significantly higher than their counterparts, compare for example Tables 5.14 and 5.27. The values of k_R is a little higher for the alternative parameters, but even for the comparable values (taking for example $k_R = 128$ in the normal case and $k_R = 2\pi * 16$ in the **ALT** case) the iteration counts are considerably higher in the **ALT** case. As nothing has changed apart from the PML parameters, it is likely that for these parameters the PMLs are performing suboptimally and may be allowing some small level of reflections (though the solutions look similar), that make the problems considerably harder to solve and cause the higher iteration counts.

4) Small improvements due to absorption occasionally, more common in ALT case

Out of the all the experiments that do not have absorption in the problem **PNA** and are not the alternative PML parameters **ALT** case, there is only one case where an improvement due to absorption is seen; in Table 5.14, $R = 3$, $k_R = 512$, for the smallest amount of absorption, the iteration count is one lower. However, in the case with the alternative PML parameters **ALT**, in Tables 5.27-5.30, in many cases with $k_R = 512$, some improvement due to absorption is seen, though the most significant ones are for $R = 3$ (see especially Table 5.28), that is not in the stable range of R for this variant of the preconditioner (see **1)**). Overall, the fact that more improvements due to absorption are seen in the **ALT** case, tends to suggest that improvements due to absorption are more likely to be observed when the problem is harder to solve (as the **ALT** case is, see **3)**) and the preconditioner exhibits higher iteration counts. This observation is born out by the experiments conducted by Shanks and his similar observation in [101, §5.4.1]. Improvements due to absorption in terms of actually reducing the iteration counts are seen in [101, Tables 5.7-10], where the experiments were conducted with wavespeed drawn from the Marmousi model (see Figure 4-12), a significantly harder problem to solve, that results in significantly higher iteration counts. (Iteration counts for increasing wavenumbers $\{70, 79, 91\}$ were reduced to $\{60, 69, 69\}$ respectively when $k_I = 1$, numbers drawn from [101, Tables 5.7 and

5.10].) This reduction in the iteration counts is in contrast to Tables [101, Tables 5.1-6] with simpler wavespeed models, where little reduction is seen in any of the iteration counts when absorption was included, similarly to our results.

5.1.4 Strongly Admissible $\mathcal{H}$-Matrix version

The algorithm for this preconditioner is exactly as in §5.1.3, recall Algorithms 5.1.5 and Algorithm 5.1.6, with the exception that we use strongly admissible $\mathcal{H}$-matrices instead of weakly admissible $\mathcal{H}$-matrices with functions hinv, hsub, hdiagnul, and hmatvec adapted accordingly. The code used for our experiments is based on the code provided by Lexing Ying that we used in §5.1.3, but we replaced the weakly admissible approximation by an adapted strongly admissible $\mathcal{H}$-matrix functionality. The code for the strongly admissible $\mathcal{H}$-matrix functions was provided by Stephanie Meier-Rohr, for some details about the exact functions used see Appendix A. This code does not use the full $\mathcal{H}$-matrix algebra, but the results do show the power of $\mathcal{H}$-matrices in building these preconditioners.

We note that this code can be viewed as a different implementation of the preconditioner used for Engquist and Ying's numerical experiments in [34] as they used strongly admissible $\mathcal{H}$-matrices. Our experiments focus on examining the effect of adding absorption (not covered in [34]) and consider a variety of different ranks R for the low-rank approximations.

5.1.4.1 Interpretation of strongly admissible preconditioner results

We discuss several aspects of the strongly admissible preconditioner results in turn.

1) Reduction in iteration counts as R increases

2) Small change in iteration counts for increased leaf-size

3) Same or higher iteration counts than in weakly-admissible case

Comparing Tables 5.31 and 5.33 we also note that we again observe that there is little change in the iteration counts when finer grids are used, apart from occasional small increases for the finer grids and that this was commented upon in 3) in §5.1.2.4.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 1$								
64	5	5	6	7	7	8	9	10
128	5	6	6	7	8	8	9	11
256	6	6	7	7	8	9	10	13
$R = 2$								
64	4	5	6	7	7	8	9	10
128	5	6	6	7	7	8	9	11
256	5	6	6	7	8	9	9	12
$R = 3$								
64	4	5	6	7	7	8	9	10
128	5	6	6	7	7	8	9	11
256	5	6	6	7	8	9	9	12
$R = 5$								
64	4	5	6	7	7	8	9	10
128	5	6	6	7	7	8	9	11
256	5	6	7	7	8	9	9	12

Table 5.31: Strongly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **12**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 1$								
64	4	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	5	6	6	7	7	8	9	12
$R = 2$								
64	3	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	4	5	6	7	7	8	9	12
$R = 3$								
64	3	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	4	5	6	7	7	8	9	12
$R = 5$								
64	3	5	6	7	7	8	9	10
128	4	5	6	7	7	8	9	11
256	4	5	6	7	7	8	9	12

Table 5.32: Strongly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **24**.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$
$R = 1$								
64	5	5	6	7	7	8	9	10
128	6	7	7	7	8	9	10	12
256	8	8	8	8	8	9	10	13
$R = 2$								
64	4	5	6	7	7	8	9	10
128	6	6	6	7	8	9	9	11
256	6	7	7	7	8	9	10	12
$R = 3$								
64	4	5	6	7	7	8	9	10
128	6	6	7	7	8	9	9	11
256	6	7	7	7	8	9	10	12

Table 5.33: Strongly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS**. Absorption level: **PNA**. Wavespeed model: **C1**. Size smallest block: **16**. For $k = 128$, $N = 512$.

k_R & R \ k_I	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$R = 1$											
64	5	4	4	4	4	4	4	4	3	3	2
128	5	5	5	5	5	5	5	5	4	3	2
256	6	6	6	6	6	6	6	5	4	3	2
$R = 2$											
64	4	4	4	4	4	4	4	4	3	3	2
128	5	5	5	5	5	4	4	4	3	3	2
256	5	5	5	5	5	5	5	5	4	3	2
$R = 3$											
64	4	4	4	4	4	4	4	4	3	3	2
128	5	5	5	5	5	4	4	4	3	3	2
256	5	5	5	5	5	5	5	5	4	3	2
$R = 5$											
64	4	4	4	4	4	4	4	4	3	3	2
128	5	5	5	5	5	4	4	4	3	3	2
256	5	5	5	5	5	5	5	5	4	3	2

Table 5.34: Strongly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1**. Source: **FPS**. Absorption level: **PWA**. Wavespeed model: **C1**. Size smallest block: **12**.

	0	0.25	0.5	0.75	1	1.5	2	$k_R^{0.25}$	$k_R^{0.5}$	$k_R^{0.75}$	$k_R^{0.9}$
$R = 1$											
64	5	5	5	4	4	4	4	4	3	3	2
128	6	6	6	6	6	6	6	6	4	3	3
256	8	8	8	7	7	7	7	7	5	3	3
$R = 2$											
64	4	4	4	4	4	4	4	4	3	3	2
128	6	5	5	5	5	5	5	5	4	3	3
256	6	6	6	6	6	6	6	5	4	3	3
$R = 3$											
64	4	4	4	4	4	4	4	4	3	3	2
128	6	6	5	5	5	5	5	5	4	3	3
256	6	6	6	6	6	6	6	6	4	3	3

Table 5.35: Strongly admissible preconditioner iteration counts. Dependence of h on k_R: **HK1.5**. Source: **FPS**. Absorption level: **PWA**. Wavespeed model: **C1**. Size smallest block: **16**.

1) Reduction in iteration counts as R increases

All the tables show reduced iteration counts as R increases, as in the weakly admissible $\mathcal{H}$-matrix preconditioner. We are able to include results for $R = 1$ and $R = 2$ as the iteration counts are stable for even these small ranks. This is because the strongly admissible condition is more restrictive or stronger, as discussed in §4.2. Some of the low-rank properties predicted for strongly admissible $\mathcal{H}$-matrices in §4.2 are seen also for weakly admissible matrix blocks in §4.3, yet the fact that the preconditioner behaving more stably in this strongly admissible case lends some weight to the idea that the low-rank theory is better justified in the strongly admissible case we use. (Engquist and Ying do find low-rank results for an off-diagonal block [34] of a weakly admissible $\mathcal{H}$-matrix, see Theorem 2.2.23, but it is unclear how justified the scaling of this result to smaller diagonal blocks is.

2) Small change in iteration counts for increased leaf-size

Tables 5.31 and 5.32 show that for leaf sizes **12** and **24** respectively, there is some small change in the iteration counts, a different effect to in the weakly admissible matrix case where there was little change for different leaf-sizes. As the leaf-size **12** is the case where more of the matrix is stored in low-rank off-diagonal blocks (i.e. approximated rather than being stored exactly, see §4.2.2 and §4.2.3), it makes sense that this should be the table with higher iteration counts, as is observed.

3) Same or higher iteration counts than in weakly-admissible case

Comparing Tables 5.14 and 5.31 for $R = 3$ and $R = 5$ show that the iteration counts in the strongly admissible version are similar or higher than the iteration counts in the weakly admissible version. This is unexpected as the strongly-admissible $\mathcal{H}$-matrices provide a better approximation to the original matrices, see §4.2. However, the iteration counts seen in the original strongly-admissible implementation of the preconditioner by Engquist and Ying are lower than either of the sets of iteration counts we see in this thesis [34, Table 3.1], so the large strongly-admissible iterations counts that we see may be due to the particular implementation of the strongly-admissible preconditioner, rather than a definitive property of the strongly-admissible preconditioner.

Chapter 6

Summary of Results

In Chapter 2 we look at previous work on sweeping preconditioners for Helmholtz problems. We recollect a particular formulation of a sweeping preconditioner and identify the importance of approximating Schur complement matrices $\mathbb{S}_m^{-1}$ (see Definition 2.2.5) that arise in the formulation. We find that the Schur complement matrices are approximations to matrices $\mathbb{G}^m$ (see Definition 2.2.7). These matrices $\mathbb{G}^m$ are formed of point-wise evaluations of a sequence of Green's functions G^m (see Definition 2.2.6) for half-plane Helmholtz problems. The Green's functions G^m are the sum of two Hankel functions (the fundamental solution of the Helmholtz equation in 2D, see §1.1.1.1, especially (1.3)). There exists low-rank approximation theory for the Hankel functions and thence the Green's functions G^m. This motivates low-rank $\mathcal{H}$-matrix approximation (see description of $\mathcal{H}$-matrices in §1.8.3) of $\mathbb{G}^m$ and thence of $\mathbb{S}_m^{-1}$.

In Chapter 3 we present our new results about the existence of low-rank separable expansions for the Hankel function. The main results are Theorems 3.2.3 and 3.2.9 (with important associated Remarks 3.2.5, 3.2.7, 3.2.10 and 3.2.11 and Lemma 3.2.8). We then use these to obtain results about the existence of low-rank separable expansions for the sequence of half-plane Green's functions G^m from Chapter 2. The main results are Theorems 3.3.3 and 3.3.7 (with important associated Remarks 3.3.4, 3.3.5 and 3.3.6 and Remarks 3.3.9 and 3.3.10 respectively). Our results differ from the previous work on low-rank results recollected in Chapter 2 because we focus here on examining the effect of absorption. In practice absorption is added to sweeping preconditioners, but little work has previously been done on the low-rank theory in the case of added absorption. We

show that for a special form of absorption, either

- a lower rank may be needed to get the same quality of approximation (Theorems 3.2.3 and 3.3.3),
- the approximation quality improves (Theorems 3.2.9 and 3.3.7),
- or that, under certain conditions, the low-rank expansions exist on much taller domains when absorption is added (Remarks 3.2.7 and 3.3.5).

In Chapter 4 we use the low-rank results for the sequence of Green's functions G^m obtained in Chapter 3, to prove new results about $\mathcal{H}$-matrix approximations to $\mathbb{G}^m$. The main results are Theorems 4.2.29, 4.2.33 and 4.2.35. We show that for a special form of absorption, either

- a lower rank may be needed to get the same quality of approximation for some $\mathcal{H}$-matrix blocks (Theorems 4.2.29),
- the approximation quality improves (Theorems 4.2.33 and 4.2.35).

Since the Schur complement matrices $\mathbb{S}_m^{-1}$ are approximately equal to the matrices $\mathbb{G}^m$ we conduct numerical experiments on $\mathcal{H}$-matrix blocks of both $\mathbb{S}_m^{-1}$ and $\mathbb{G}^m$. Here we do empirical studies including the case of variable wavespeed $c(x)$ in $\mathbb{S}_m^{-1}$. This is not covered in the Green's function theory, which requires constant wavespeed, see §1.1.1. We find very good correlation between the results of the numerical experiments and the theoretical results, observing many properties we expect to find in consequence of the improvements due to absorption shown in the theory, see the analysis in §4.3.3-4.3.6. A summary of the analysis is as follows:

- the dependence of the rank upon the quality of the approximation (Experiment 1 in §4.3.3),
- the independence of the rank on the wavenumber (Experiment 2 in §4.3.4),
- the exponential improvement in the quality of the approximation with absorption (Experiment 3 in §4.3.5),
- and the ranks still being low for fatter domains when absorption is included (Experiment 4 in §4.3.6).

In Chapter 5 we perform experiments with existing sweeping preconditioners – both in the common scenario when the problem to be solved does not contain absorption (see Definition 1.9.4) and also the less common scenario when the problem to be solved does contain absorption. These experiments are novel in that they particularly study the effect of adding absorption. Specifically our goal is to investigate whether the benefits due to absorption seen in the $\mathcal{H}$-matrix approximations of the Schur complement matrices $\mathbb{S}_m^{-1}$ translate to benefits in the performance of the preconditioner. We also consider the effect of varying the rank R in the $\mathcal{H}$-matrix approximation of the Schur complements and the height of quasi-1D problems solved as part of the moving PML method (see §5.1.2). The main analysis of the results is summarised in §5.1.2.5, §5.1.3.2 and §5.1.4.1. In some cases we see improvements due to absorption and in others we do not. The performance of the iterative method is highly dependent on the parameters used in both the discretisation of the problem and the construction of the preconditioners. An example of the former is **4)** in §5.1.3.2, that says that small improvements due to absorption more common in the **ALT** case (see §5.1.1.6) where the PML has different parameters. An example of the latter is **4)** in §5.1.2.5, that says that improvement due to absorption are more likely to be seen for smaller amounts of absorption and lower values of D.

Appendix A

The Strongly Admissible $\mathcal{H}$-matrix Functions

Here we give details about the functions hsub, hdiagmul, hmatvec and hinv, coded by Stephanie Meier-Rohr and used in the strongly admissible $\mathcal{H}$-matrix preconditioner in §5.1.4. We give details only at the level of a sub-block of a $\mathcal{H}$-matrix, where the result should also be a sub-block of a $\mathcal{H}$-matrix.

1) hsub To add two $\mathcal{H}$-matrix blocks $A = UV^T$ and $B = U'V'^T$, first the low-rank matrix factors are concatenated as follows: $A' = [U, U']$, $B' = [V, V']^T$. Then QR-decompositions of the concatenations are found $[Q, T] = \text{qr}(A')$, $[Q', T'] = \text{qr}(B')$. Then the SVD of the multiple of the two 'R' matrices (T and T') is found as follows: $[U'', \Sigma'', V''] = \text{svd}(TT')$. Finally the two low-rank matrices are created as follows: calculate $U''' = QU''\Sigma''$ and truncate it to the first R columns and calculate $V''' = Q'V''$ and truncate it to the first R rows.

2) hdiagmul This operation multiplies a $\mathcal{H}$-matrix and a diagonal matrix, but is not done taking into account the fact that the diagonal matrix is sparse; instead it is computed as a normal $\mathcal{H}$-matrix multiplication. The normal $\mathcal{H}$-matrix multiplication is as follows.

To multiply the two $\mathcal{H}$-matrix blocks $A = UV^T$ and $B = U'V'^T$, their multiplication is given by $AB = U(V^TU')V'^T = U(V'U'^TV)^T$, which is already in the low-rank form of a $\mathcal{H}$-matrix block.

3) hmatvec To multiply the $\mathcal{H}$-matrix block $A = UV^T$ by the vector x, first compute $y = V^T x$ and then Uy.

4) hinv calculates a full-rank inverse of the low-rank matrix according to [57, Algorithm 5] and compresses it to a low-rank $\mathcal{H}$-matrix using a truncation of the SVD. (The truncation of the SVD is as follows: for a low-rank matrix of rank R, the appropriate first R singular vectors of U' and $\Sigma V'^*$ (from the SVD in Theorem 4.3.1) form the low-rank approximation for each block).

Bibliography

[1] *The Holy Bible, English Standard Version*. Crossway, 2001.

[2] M. Ablowitz and A. Fokas. *Complex Variables: Introduction and Applications*. Cambridge University Press, second edition, 2003.

[3] L. Banjai. Multistep and Multistage Convolution Quadrature for the Wave Equation: Algorithms and Experiments. *SIAM J. Sci. Comput.*, 32(5):2964–2994, 2010.

[4] R. Barrett, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, Philadelphia, 1994.

[5] A. Bayliss, C. I. Goldstein, and E. Turkel. On accuracy conditions for the numerical computation of waves. *Journal of Computational Physics*, 59:396–404, 1985.

[6] M. Bebendorf. *Hierarchical matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems*, volume 63 of *Lecture Notes in Computational Science and Engineering*. Springer, 2008.

[7] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov. Some Remarks on the Elman Estimate for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 27(3):772–778, 2006.

[8] J.-P. Berenger. A perfectly matched layer for the absorption of electromagnetic waves. *Journal of Computational Physics*, 114(2):185–200, 1994.

[9] M. A. Biot. Theory of Propagation of Elastic Waves in a Fluid-Saturated Porous Solid. II. Higher Frequency Range. *The Journal of the Acoustical Society of America*, 28(2):179–191, 1956.

[10] S. Börm, M. Löhndorf, and J. M. Melenk. Approximation of Integral Operators by Variable-Order Interpolation. *Numerische Mathematik*, 99(4):605–643, 2005.

[11] S. Börm and J. M. Melenk. Approximation of the high-frequency Helmholtz kernel by nested directional interpolation: error analysis. *Numerische Mathematik*, 137(1):1–34, 2017.

[12] P. Cance and Y. Capdeville. Validity of the acoustic approximation for elastic waves in heterogeneous media. *Geophysics*, 80(4):T161–T173, 2015.

[13] S. N. Chandler-Wilde, I. G. Graham, S. Langdon, and E. A. Spence. Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering. *Acta Numerica*, 21:89–305, 2012.

[14] S. Chandrasekaran, M. Gu, and T. Pals. A Fast ULV Decomposition Solver for Hierarchically Semiseparable Representations. *SIAM Journal on Matrix Analysis and Applications*, 28(3):603–622, 2006.

[15] Z. Chen and X. Xiang. A Source Transfer Domain Decomposition Method for Helmholtz Equations in Unbounded Domain. *SIAM Journal on Numerical Analysis*, 51(4):2331–2356, 2013.

[16] W. C. Chew, J. M. Song, T. J. Cui, S. Velamparambil, M. L. Hastriter, and B. Hu. Review of Large Scale Computing in Electromagnetics with Fast Integral Equation Solvers. *CMES - Computer Modeling in Engineering and Sciences*, 5(4):361–372, 2004.

[17] P.-H. Cocquet and M. J. Gander. On the Minimal Shift in the Shifted Laplacian Preconditioner for Multigrid to Work. In T. Dickopf, M. J. Gander, L. Halpern, R. Krause, and L. F. Pavarino, editors, *Domain Decomposition Methods in Science and Engineering XXII*, volume 104 of *Lecture Notes in Computational Science and Engineering*, pages 137–145. Springer, 2016.

[18] P.-H. Cocquet and M. J. Gander. How Large a Shift is Needed in the Shifted Helmholtz Preconditioner for its Effective Inversion by Multigrid? *SIAM J. Sci. Comput.*, 39(2):A438–A478, 2017.

[19] P.-H. Cocquet and M. J. Gander. Analysis of the Shifted Helmholtz Expansion Preconditioner for the Helmholtz Equation. In P. Bjørstad, S. Brenner, L. Halpern, H. Kim, R. Kornhuber, T. Rahman, and O. Widlund, editors, *Domain Decomposition Methods in Science and Engineering XXIV, Conference proceedings DD 2017*, volume 125 of *Lecture Notes in Computational Science and Engineering*, pages 195–203. Springer, Cham, 2018.

[20] F. Collino and C. Tsogka. Application of the perfectly matched absorbing layer model to the linear elastodynamic problem in anisotropic heterogeneous media. *Geophysics*, 66(1):294–307, 2001.

[21] D. Colton and R. Kress. *Integral equation methods in scattering theory.* Wiley, 1983.

[22] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*, volume 93 of *Applied Mathematical Sciences.* Springer, 3rd edition, 2013.

[23] S. Cools and W. Vanroose. Local Fourier analysis of the complex shifted Laplacian preconditioner for Helmholtz problems. *Numer. Linear Algebra Appl.*, 20:575–597, 2013.

[24] W. Dahmen, S. Prössdorf, and R. Schneider. Wavelet approximation methods for pseudodifferential equations II: Matrix compression and fast solution. *Advances in Computational Mathematics*, 1(3):259–335, 1993.

[25] K. Delamotte. *Une étude du rang du noyau de l'équation de Helmholtz: application des $\mathcal{H}$-matrices à l'EFIE.* PhD thesis, University Paris 13, 2016.

[26] K. Delamotte, T. Abboud, and O. Lafitte. On the Local Approximate Rank of Helmholtz Green's Kernel. In *WAVES 2017, Minneapolis*, pages 103–4, 2017.

[27] R. A. DeVore and G. G. Lorentz. *Constructive approximation.* Springer-Verlag, 1993.

[28] G. C. Diwan, A. Moiola, and E. A. Spence. Can coercive formulations lead to fast and accurate solution of the Helmholtz equation? *arXiv:1806.05934*, 2018.

[29] J. Douglas, J. E. Santos, D. Sheen, and L. S. Bennethum. Frequency Domain Treatment of One-Dimensional Scalar Waves. *Mathematical models and methods in applied sciences*, 3(2):171–194, 1993.

[30] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational Iterative Methods for Nonsymmetric Systems of Linear Equations. *SIAM Journal on Numerical Analysis*, 20(2):345–357, 1983.

[31] H. C. Elman. *Iterative methods for large, sparse, nonsymmetric systems of linear equations*. PhD thesis, Yale University, 1982.

[32] B. Engquist and A. Majda. Absorbing boundary conditions for numerical simulation of waves. *Proceedings of the National Academy of Sciences*, 74(5):1765–1766, 1977.

[33] B. Engquist and L. Ying. Fast Directional Multilevel Algorithms for Oscillatory Kernels. *SIAM Journal on Scientific Computing*, 29(4):1710–1737, 2007.

[34] B. Engquist and L. Ying. Sweeping preconditioner for the Helmholtz equation: Hierarchical matrix representation. *Communications on Pure and Applied Mathematics*, 64(5):697–735, 2011.

[35] B. Engquist and L. Ying. Sweeping Preconditioner for the Helmholtz Equation: Moving Perfectly Matched Layers. *Multiscale Modeling & Simulation*, 9(2):686–710, 2011.

[36] B. Engquist and H. Zhao. Approximate Separability of Green's Function for High Frequency Helmholtz Equations. Technical report, University of California, Los Angeles, 2014.

[37] Y. Erlangga, C. Vuik, and C. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50(3):409–425, 2004.

[38] O. Ernst and M. Gander. Why it is Difficult to Solve Helmholtz Problems with Classical Iterative Methods. In I. G. Graham, R. Scheichl, O. Lakkis, and T. Y. Hou, editors, *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 325–363. Springer Berlin Heidelberg, 2012.

[39] R. Fletcher. Conjugate Gradient Methods for Indefinite Systems. In G. A. Watson, editor, *Numerical Analysis. Lecture Notes in Mathematics*, volume 506. Springer, Berlin Heidelberg, 1976.

[40] A. Frangi and M. Bonnet. On the application of the fast multipole method to helmholtz-like problems with complex wavenumber. *CMES - Computer Modeling in Engineering and Sciences*, 58(3):271–296, 2010.

[41] T. C. Frelet. *Finite element approximation of Helmholtz problems with application to seismic wave propagation.* PhD thesis, INSA de Rouen, 2015.

[42] M. J. Gander, I. G. Graham, and E. A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed? *Numerische Mathematik*, 131(3):567–614, 2015.

[43] M. J. Gander and S. Solovyev. A Numerical Study on the Compressibility of Subblocks of Schur Complement Matrices Obtained from Discretized Helmholtz Equations. In I. Dimov, I. Faragó, and L. Vulkov, editors, *Numerical Analysis and Its Applications. NAA 2016.*, pages 70–81. Spinger, Cham, 2017.

[44] M. J. Gander and H. Zhang. Iterative Solvers for the Helmholtz Equation: Factorizations, Sweeping Preconditioners, Source Transfer, Single Layer Potentials, Polarized Traces, and Optimized Schwarz Methods. *arXiv:1610.02270 [math.NA]*, 2016.

[45] M. J. Gander and H. Zhang. Optimized Schwarz Methods with Overlap for the Helmholtz Equation. 38(5):A3195–A3219, 2016.

[46] P. Gatto and J. Hesthaven. A Preconditioner Based on Low-Rank Approximation of Schur Complements. *arXiv:1508.07798v1 [math.NA]*, 2015.

[47] P. Gerhard, C. Shin, and Hicks. Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophysical Journal International*, 133(2):341–362, 1998.

[48] A. Gillman, A. H. Barnett, and P. G. Martinsson. A spectrally accurate direct solution technique for frequency-domain scattering problems with variable media. *BIT Numerical Mathematics*, 55(1):141–170, 2015.

[49] A. Gillman and P. G. Martinsson. An O(N) algorithm for constructing the solution operator to 2D elliptic boundary value problems in the absence of body loads. *Advances in Computational Mathematics*, 40(4):773–796, 2014.

[50] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, second edition, 1989.

[51] I. G. Graham, W. Hackbusch, and S. A. Sauter. Finite elements on degenerate meshes: inverse-type inequalities and applications. *IMA Journal of Numerical Analysis*, 25(2):379–407, 2005.

[52] I. G. Graham, M. Löhndorf, J. M. Melenk, and E. A. Spence. When is the error in the h–BEM for solving the Helmholtz equation bounded independently of k? *BIT Numerical Mathematics*, 55:171–214, 2015.

[53] I. G. Graham, E. A. Spence, and E. Vainikko. Domain Decomposition Preconditioning for High-Frequency Helmholtz Problems with Absorption. *Mathematics of Computation*, 86(307):2089–2127, 2017.

[54] I. G. Graham, E. A. Spence, and E. Vainikko. Recent Results on Domain Decomposition Preconditioning for the High-Frequency Helmholtz Equation Using Absorption. In D. Lahaye, J. Tang, and K. Vuik, editors, *Modern Solvers for Hemholtz Problems*, Geosystems Mathematics, pages 3–26. Springer International Publishing, 2017.

[55] I. G. Graham, E. A. Spence, and J. Zou. Domain Decomposition with local impedance conditions for the Helmholtz equation. *arXiv:806.03731v2*, 2018.

[56] L. Grasedyck and W. Hackbusch. Construction and arithmetics of $\mathcal{H}$-matrices. *Computing*, 70(4):295–334, 2003.

[57] L. Grasedyck, W. Hackbusch, and S. Börm. *Hierarchical Matrices: lecture note no.21, 2003. Revised 2006.* Max Planck Institut für Mathematik, Bonn, available at http://www.mis.mpg.de/preprints/ln/lecturenote-2103.pdf, accessed on 2015-12-15.

[58] A. Greenbaum. *Iterative Methods for Solving Linear Systems*, volume 17 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 1997.

[59] A. Greenbaum, V. Pták, and Z. Strakoš. Any Nonincreasing Convergence Curve is Possible for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 17(3):465–469, 1996.

[60] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348, 1987.

[61] W. Hackbusch. A Sparse Matrix Arithmetic Based on $\mathcal{H}$-Matrices. Part I: Introduction to $\mathcal{H}$-Matrices. *Computing*, 62(2):89–108, 1999.

[62] W. Hackbusch. *Multi-grid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer Science Business Media, 2003.

[63] W. Hackbusch. *Hierarchical Matrices: Algorithms and Analysis*, volume 49 of *Springer Series in Computational Mathematics*. 2015.

[64] W. Hackbusch and S. Börm. Data-sparse approximation by adaptive $\mathcal{H}^2$-matrices. *Computing*, 69(1):1–35, 2002.

[65] W. Hackbusch and Z. P. Nowak. On the fast matrix multiplication in the boundary element method by panel clustering. *Numerische Mathematik*, 54(4):463–491, 1989.

[66] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011.

[67] M. R. Hestenes and E. Stiefel. Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

[68] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 1996.

[69] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer Verlag, New York, 1998.

[70] F. Ihlenburg and I. Babuška. Dispersion Analysis and Error Estimation of Galerkin Finite Element Methods for the Helmholtz Equation. *International Journal for Numerical Methods in Engineering*, 38(22):3745–3774, 1995.

[71] F. Ihlenburg and I. Babuška. Finite Element Solution of the Helmholtz Equation with High Wave Number Part II: the $h - p$ Version of the FEM. *SIAM Journal on Numerical Analysis*, 34(1):315–358, 1997.

[72] F. Ihlenburg and I. M. Babuška. Finite Element Solution of the Helmholtz Equation with High Wave Number Part I: The h-Version of the FEM. *Comput. Math. Appl.*, 30(9):9–37, 1995.

[73] M. Kachanovska. Hierarchical matrices and the High-Frequency Fast Multipole Method for the Helmholtz Equation with Decay. *Max Planck Institute, Leipzig preprint no.13*, (March), 2014.

[74] C. T. Kelly. *Iterative Methods for Linear and Nonlinear Equations*, volume 16 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 1995.

[75] T. Kim. *Asymptotic and Numerical methods for high-frequency scattering problems*. PhD thesis, University of Bath, 2012.

[76] S. Kobayashi and N. Nishimura, editors. *Boundary Element Methods: Fundamentals and Applications*. Springer Science and Business Media, 2013.

[77] W. Y. Kong, J. Bremer, and V. Rokhlin. An adaptive fast direct solver for boundary integral equations in two dimensions. *Applied and Computational Harmonic Analysis*, 31(3):346–369, 2011.

[78] P. Lailly. *The seismic inverse problem as a sequence of before stack migrations*. Conference on Inverse Scattering: Theory and Application. SIAM, Philadelphia, 1983.

[79] G. Leoni. *A First Course in Sobolev Spaces*. American Mathematical Society, Providence, 2nd edition, 2017.

[80] F. Liu and L. Ying. Additive Sweeping Preconditioner for the Helmholtz Equation. *Multiscale Modeling & Simulation*, 14(2):799–822, 2016.

[81] P. G. Martinsson and V. Rokhlin. A fast direct solver for boundary integral equations in two dimensions. *Journal of Computational Physics*, 205(1):1–23, 2005.

[82] P. G. Martinsson and V. Rokhlin. A fast direct solver for scattering problems involving elongated structures. *Journal of Computational Physics*, 221(1):288–302, 2007.

[83] J. M. Melenk. *On Generalised Finite Element Methods*. PhD thesis, The University of Maryland, 1995.

[84] J. M. Melenk and S. A. Sauter. Wavenumber Explicit Convergence Analysis for Galerkin Discretisations of the Helmholtz Equation. *SIAM Journal on Numerical Analysis*, 49(3):1210–1243, 2011.

[85] L. Métivier, R. Brossier, J. Virieux, and S. Operto. Full Waveform Inversion and the Truncated Newton Method. *SIAM Review*, 59(1):153–195, 2017.

[86] A. Moiola and E. A. Spence. Is the Helmholtz Equation Really Sign-Indefinite? *SIAM Review*, 56(2):274–312, 2014.

[87] Nwhit. Image used under CC BY-SA 3.0, http://creativecommons.org/licenses/by-sa/3.0, edited to simplify and remove text https://commons.wikimedia.org/wiki/File:Diagram_of_a_marine_seismic_survey.png. Accessed on 05/05/17.

[88] P. Ocłoń, S. Łopata, and M. Nowak. Comparative study of conjugate gradient algorithms performance on the example of steady-state axisymmetric heat transfer problem. *Archives of Thermodynamics*, 34(3):15–44, 2013.

[89] F. W. J. Olver. Bessel Functions of Integral Order. In M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Math-*

ematics Series, chapter 9, pages p355–434. National Bureau of Standards. Accessed http://people.math.sfu.ca/∼cbm/aands/page_358.htm on 27/04/2019, 1972.

[90] F. W. J. Olver. *Asymptotics and Special Functions*. Computer Science and Applied Mathematics. Academic Press, New York, 1974.

[91] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions. Print companion to [92]*. Cambridge University Press, New York, 2010.

[92] F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. V. Saunders. *NIST Digital Library of Mathematical Functions. Online companion to [91]*. http://dlmf.nist.gov/, Release 1.0.10 of 2015-08-07.

[93] F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. V. Saunders. *NIST Digital Library of Mathematical Functions. Online companion to [91]*. http://dlmf.nist.gov/ Release 1.0.22 of 2019-03-15.

[94] C. C. Paige and M. A. Saunders. Solution of Sparse Indefinite Systems of Linear Equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[95] R. G. Pratt. Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model. *Geophysics*, 64(3):888–901, 1999.

[96] R. G. Pratt, W. J. McGaughey, and C. H. Chapman. Anisotropic velocity tomography: A case study in a near-surface rock mass. *Geophysics*, 58(12):1748–1763, 1993.

[97] H. A. Priestley. *Introduction to Complex Analysis*. Oxford University Press, Oxford, revised edition, 1990.

[98] T. J. Rivlin. *The Chebyshev Polynomials*. Wiley-Interscience, New York, 1990.

[99] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, 2nd edition, 2003.

[100] Y. Saad and M. H. Schultz. GMRES: a Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM J. Sci. Stat. Comput.*, 7(3):856–869, 1986.

[101] J. D. Shanks. *Robust solvers for large indefinite systems in seismic inversion*. PhD thesis, University of Bath, 2014.

[102] R. A. Silverman. *Complex Analysis with Applications*. Dover Publications, Mineola, 1974.

[103] I. Singer and E. Turkel. High-order finite difference methods for the Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 163(1-4):343–358, 1998.

[104] P. Sonneveld. CGS, A Fast Lanczos-Type Solver for Nonsymmetric Linear Systems. *SIAM J. Sci. Stat. Comput.*, 10(1):36–52, 1989.

[105] E. A. Spence. "When all else fails, integrate by parts" - an overview of new and old variational formulations for linear elliptic PDEs. In A. Fokas and B. Pelloni, editors, *Unified Transform Method for Boundary Value Problems: Applications and Advances*. SIAM, 2015.

[106] C. Stolk. A rapidly converging domain decomposition method for the Helmholtz equation. *Journal of Computational Physics*, 241:240–252, 2013.

[107] A. Tarantola. Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8):1259–1266, 1984.

[108] L. H. Thomas. Elliptic problems in linear difference equations over a network. *Watson Sci. Comput. Lab. Rept., Columbia University, New York*, 1949.

[109] E. C. Titchmarsh. *The Theory of Functions*. Oxford University Press, Oxford, 1932.

[110] H. A. van der Vorst. Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems. *SIAM J. Sci. Stat. Comput.*, 13(2):631–644, 1992.

[111] H. A. van der Vorst. *Iterative Krylov Methods for Large Linear Systems*, volume 13. Cambridge University Press, Cambridge, 2003.

[112] R. J. Versteeg. The Marmousi experience: Velocity model determination on a synthetic complex data set. *The Leading Edge*, 13(9):927–936, 1994.

[113] H. Wu. Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. Part I: Linear version. *IMA Journal of Numerical Analysis*, 34(3):1266–1288, 2014.

[114] L. Ying and F. Liu. Recursive Sweeping Preconditioner for the Three-Dimensional Helmholtz Equation. *SIAM J. Sci. Comput.*, 38(2):A814–A832, 2013.

[115] L. Zepeda-Núñez and L. Demanet. The method of polarized traces for the 2D Helmholtz equation. *Journal of Computational Physics*, 308:347–388, 2016.

[116] O. Zienkiewicz, R. Taylor, and J. Zhu. *Finite Element Method - Its Basis and Fundamentals*. Elsevier Butterworth-Heinemann, 6th edition, 2005.